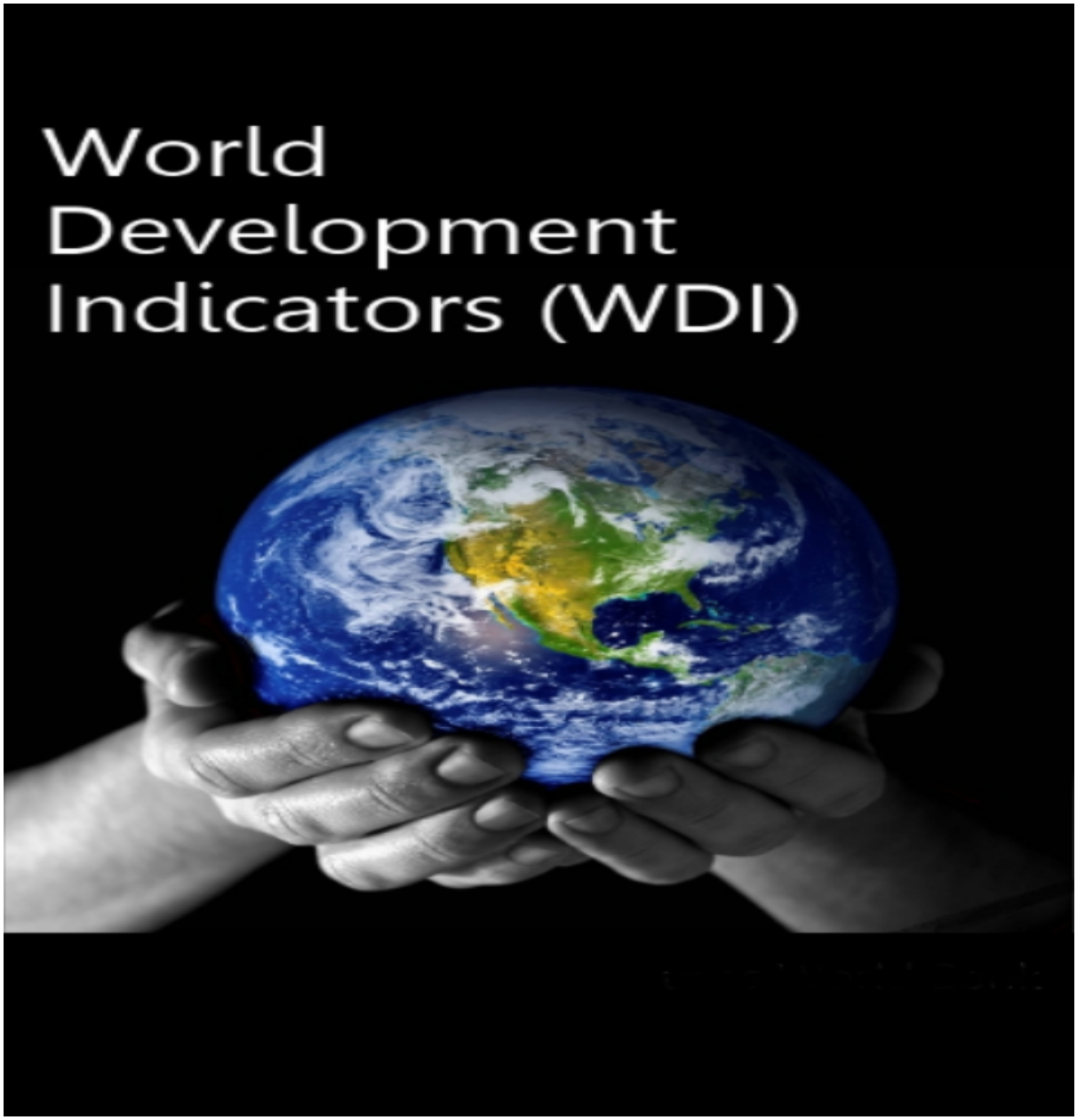
Please Note: This is a sample project report for your reference purposes. It is by no means a definitive document and you are not bound to prepare their project reports in the same manner. You are free to use any other style of presentation. This report is just for your reference.



Abstract

The primary World Bank collection of development indicators, compiled from officially-recognized international sources.It presents the most current and accurate global development data available, and includes national, regional and global estimates. It provides high-quality cross-country comparable statistics about development and people's lives around the globe. From this dataset through top-n and comparative analysis we have highlighted the developed countries data;and those countries which are in comparatively lower level to plan and take necessary step whichaims to produce a curate set of indicators relevant to the changing needs of the development community to improve thequality of life.

Project Introduction

World Development Indicator - World Bank's compilation of internationally comparable statistics about global development and the quality of people's lives. It contains data about Agriculture & Rural Development, Aid Effectiveness, Climate Change, Economy & Growth, Education, Energy & Mining, Environment, External Debt, Financial Sector, Gender, Health, Infrastructure, Labour& Social Protection, Poverty, Private Sector, Public Sector, Science & Technology, Social Development, Trade, Urban Development.

Sharing knowledge will be crucial to end extreme poverty and boost shared prosperity around the world. The World Bank Group's Global Practices bring together knowledge and expertise in 14 sectors and 5 cross-cutting areas. The goal is to help developing countries find solutions to the toughest global and local development challenges—from adapting to climate change to boosting food security or increasing access to energy.

About the dataset

The dataset consists of a single CSV file – Dataset.csv. The columns in the CSV are

- Country Name
- Country Code
- Indicator Name
- Indicator Code
- Year
- Value

File Name	Format	Size
Dataset.csv	CSV	344.41 MB

Business Questions Identified

The following questions were identified by the team after many rounds of inspection and discussions:

1. Below are the reports which will show us the top 10 countries progress in global development and quality of people's development,
    a. Top 10 countries which are exporting Arms
    b. Top 10 countries which are importing Arms
    c. Top 10 countries which are emitting CO2.
    d. Top 10 countries which are having highest Population density.
    e. Top 10 countries which are having population growth.

2. Below are the reports which will show us the Least 5 countries global development and quality of people's development
    a. Least 5 countries which are emitting low CO2.
    b. Least 5 countries with low Population between 0-14 years (% of total).
    c. Least 5 countries with low Population between 15-64 years (% of total)
    d. Least 5 countries with low Population growth.

3. Comparison report: CO2 emission report between Brazil and India.
4. Visualization of results observed

Kindly scroll down further to find the details of the approach followed by the team to answer the business questions identified.

Top 10 countries which are exporting Arms

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 10 countries which are exporting Arms to other countries. This will help the business to identify the countries which are exporting Arms

Approach

The following steps were followed to answer the business question:

- Loaded the dataset (Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT, SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Arms_exports' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'Arms exports (SIPRI trend indicator values)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2000' has been considered for extracting Top 10 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset)with SName (Arms exports (SIPRI trend indicator values)) and year (2000) and sorted the results by 'Value' column in descending order.
- Limited the sorted result set to Top 10 rows.
- For each item from that Top 10 rows generated Indicator Name(SeriesName), CountryCode, Value and Year;
- Finally the output is written into OUTPUT variable defined above in Pig program

Findings

Here are the top 10 countries which are exporting Arms

High Income	23911999500
OECD Members	23910000600
High income: OECD	23910000600
North America	15768000500
United States	15440000000
United States	8634000400

Europe and Central Asia (all income levels)	8008000000
European Union	7624000000
Euro Area	4701000200
United Kingdom	2446000130

Top 10 countries which are Importing Arms

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 10 countries which are importing Arms from other countries.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset (Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT, SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Arms_Imports' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'Arms exports (SIPRI trend indicator values)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2000' has been considered for extracting Top 10 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset)with SName (Arms Imports (SIPRI trend indicator values)) and year (2000) and sorted the results by 'Value' column in descending order.
- Limited the sorted result set to Top 10 rows.
- For each item from that Top 10 rows generated Indicator Name(SeriesName), CountryCode, Value and Year;
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the top 10 countries which are importing Arms

Europe & Central Asia (ECS)	8386999800
Middle East & North Africa (MNA)	8700999700
Upper middle income (UMC)	11779000300

Middle East & North Africa (all income levels) (MEA)	11932999700
High income: OECD (OEC)	12334999600
OECD members (OED)	12837999600
High income (HIC)	15047000100
Middle income (MIC)	15678000100
Low & middle income (LMY)	17116000300
World (WLD)	32163000300

Top 10 countries which are emitting CO2

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 10 countries which are emitting high CO2.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset (Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT, SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Co2_emission_kt' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'CO2 Emission (SIPRI trend indicator values)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2000' has been considered for extracting Top 10 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset)with SName (CO2 emission (SIPRI trend indicator values)) and year (2000) and sorted the results by 'Value' column in descending order.
- Limited the sorted result set to Top 10 rows.
- For each item from that Top 10 rows generated Indicator Name(SeriesName), CountryCode, Value and Year;
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the top 10 countries which are emitting high CO2

World	17726098.0
High income	12594944.0
OECD members	10441799.0
High income: OECD	10205053.0
East Asia & Pacific (all income levels)	6547425.0
North America	5012888.0
United States	4613100.5
European Union	4332074.0
United States	4328904.5
Low and Middle income	3803208.2

Top 10 countrieswith highest Population density (people per sq. km of land area)

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 10 countries with highest population density (calculated as people per sq. km of land area).

Approach

The following steps were followed to answer the business question:

- Loaded the dataset (Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT, SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in ' Population_density' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is ' Population_density (SIPRI trend indicator values)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program , year '2000' has been considered for extracting Top 10 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset)with SName (Population_density (SIPRI trend indicator values)) and year (2000) and sorted the results by 'Value' column in descending order.
- Limited the sorted result set to Top 10 rows.
- For each item from that Top 10 rows generated Indicator Name(SeriesName), CountryCode, Value and Year;

- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the top 10 countries with highest Population density

MCO	12760.5
SGP	3422.8537
BMU	1064
MLT	955.5438
CHI	654.1546
BRB	575.4512
BGD	560.2689
MDV	468.6067
MUS	446.5552
BHR	411.367

Top 10 countries with highest Population growth (annually)

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 10 countries with percentage growth of population annually.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset (Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT, SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in ' Population_ growth' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is ' Population_ growth (SIPRI trend indicator values)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2000' has been considered for extracting Top 10 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset)with SName (Population_growth(SIPRI trend indicator values)) and year (2000) and sorted the results by 'Value' column in descending order.
- Limited the sorted result set to Top 10 rows.

- For each item from that Top 10 rows generated Indicator Name(SeriesName), CountryCode, Value and Year;
- Finally the output is written into OUTPUT variable defined above in Pig program

Findings

Here are the top 10 countries with growth of population annually

Monaco(MCO)	12760.5
Singapore(SGP)	3422.8537
Bermuda(BMU)	1064
Malta(MLT)	955.5438
Channel Islands(CHI)	654.1546
Barbados(BRB)	575.4512
Bangladesh(BGD)	560.2689
Maldives(MDV)	468.6067
Mauritius(MUS0	446.5552
Bahrain(BHR)	411.367

Least 5 countries which are emitting lowest CO2

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the top 5 countries which are emitting lowest CO2.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT,SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Co2_emissions' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'CO2 emissions (metric tons per capita)' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2001' has been considered for extracting Least 5 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset(A_dataset)with SName (CO2 emissions (metric tons per capita)) and year (2001) and sorted the results by 'Value' column in ascending order.
- Limited the sorted result set to least 5 rows.

- For each item from that least 5 rows generated Indicator Name(SeriesName), CountryCode, Value
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the least 5 countries with lowest emission of CO2

Cambodia(KHM)	0.012
Bhutan(BTN)	3422.8537
Rwanda(RWA)	1064
Nepal(NPL)	955.5438
Bhutan (BTN)	654.1546

Least 5 countries with low Population between 0-14 years (% of total)

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the low population between 0-14 years (% of total).

Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT,SName, year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in ' Population_ages_0_14' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'Population_ages_0_14' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2001' has been considered for extracting Least 5 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset(A_dataset)with SName (Population_ages_0_14) and year (2001) and sorted the results by 'Value' column in ascending order.
- Limited the sorted result set to least 5 rows.
- For each item from that least 5 rows generated Indicator Name(SeriesName), CountryCode, Value
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the least 5 countries with low population between 0-14 years (% of total)

Italy(ITA)	12.3%
Canada(CAN)	16.3%
East Asia & Pacific (EAS)	22.0%
Argentina(ARG)	24.0%
Brazil(BRA)	25.3%

Least 5 countries with low Population between 15-64 years (% of total)

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the least 5 countries with low population between 15-64 year (% of total).

Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT,SName, and year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in ' Population_ages_15_64' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is ' Population_ages_15_64' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2001' has been considered for extracting Least 5 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset(A_dataset)with SName (Population_ages_15_64) and year (2001) and sorted the results by 'Value' column in ascending order.
- Limited the sorted result set to least 5 rows.
- For each item from that least 5 rows generated Indicator Name(SeriesName), CountryCode, Value
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the least 5 countries with low population between 15-64 year (% of total)

Afghanistan(AFG)	16.7%
Arab World(ARB)	19.7%
India(IND)	20.7%
Argentina(ARG)	21.1%
Indonesia(IDN)	21.9%

Least 5 countries with low Population growth

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the least 5 countries with low population growth.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT,SName, year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in ' Population_growth ' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'Population_growth' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program, year '2001' has been considered for extracting Least 5 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset(A_dataset)with SName (Population_growth) and year (2001) and sorted the results by 'Value' column in ascending order.
- Limited the sorted result set to least 5 rows.
- For each item from that least 5 rows generated Indicator Name(SeriesName), CountryCode, Value
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the least 5 countries with low population growth

Equatorial Guinea(GNQ)	-3.6271
Timor-Leste(TMP)	-3.1766
Virgin Islands(U.S)(VIR)	-3.1749
Cambodia(KHM)	-2.859
Equatorial Guinea(GNQ)	-2.0806

Least 5 countries which are having largest population

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the least 5 countries with population city.

Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared four variables- INPUT, OUTPUT,SName, year.
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Population_largest_city ' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'Population_largest_city ' in our case.
- Year variable holds 'Year' column from Dataset.csv. For this program , year '2001' has been considered for extracting Least 5 indicator values.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset(A_dataset)with SName ('Population_largest_city) and year (2001) and sorted the results by 'Value' column in ascending order.
- Limited the sorted result set to least 5 rows.
- For each item from that least 5 rows generated Indicator Name(SeriesName), CountryCode, Value
- Finally the output is written into OUTPUT variable defined above in Pig program.

Findings

Here are the least 5 countries with low population city

East Asia & Pacific (EAS)	0.0
Arab World(ARB)	0.1
Oman(OMN)	45288.0
Mauritania(MRT)	46077.0
South Sudan(SSD)	47855.0

Comparison report: CO2 emission report between Brazil and India

Business Question

The Dataset.csv consists data of different type's indicators by which we can explore the global development. The business wants to know the comparison of CO2 emissions between Brazil and India.

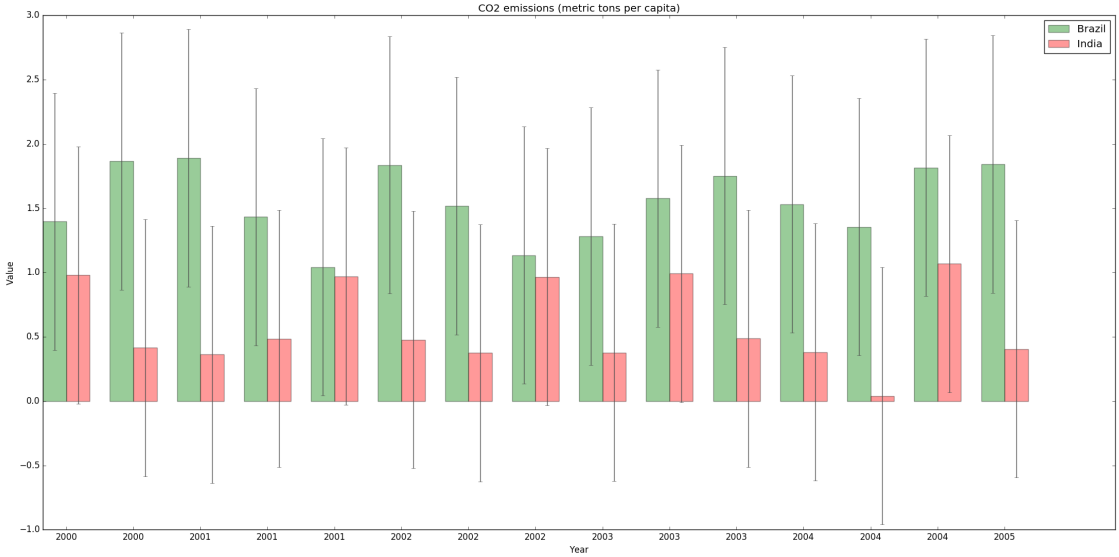
Approach

The following steps were followed to answer the business question:

- Loaded the dataset(Dataset.csv) file into HDFS.
- Created a pig program and have declared variables- INPUT, OUTPUT ,SName, FromYear,ToYear, Country1, Country2
- INPUT file referencing complete path of actual data set file(Dataset.csv)
- The output of this pig program will be stored in 'Co2_emissions' folder and OUTPUT variable holds the complete path of this output folder.
- SName variable holds the 'Indicator Name' Column from Dataset.csv file which is 'CO2 emissions (metric tons per capita)' in our case.
- FromYear variable holds 'Year' column from Dataset.csv. Year '2000' has been considered as the stating year for comparing indicator values.
- ToYear variable holds 'Year' column from Dataset.csv. For this program, Year '2005' has been considered as the end year for comparing indicator values.
- We are comparing the indicator values between two countries, Brazil and India which are being stored in Country1 and Country2 variables respectively.
- Created another variable 'A_dataset' and loaded the dataset (Dataset.csv) to this variable and defined the schema.
- Filtered the dataset (A_dataset) with Sname (CO2 emissions (metric tons per capita)) and with the years between 2001 and 2005. Also filtered the dataset with countries with Brazil and India and sorted the results by CountryCode and Year.
- For each item from result set generated CountryName , Indicator Name(SeriesName), Value and Year.
- Finally the output is written into OUTPUT variable defined above in Pig program

Findings

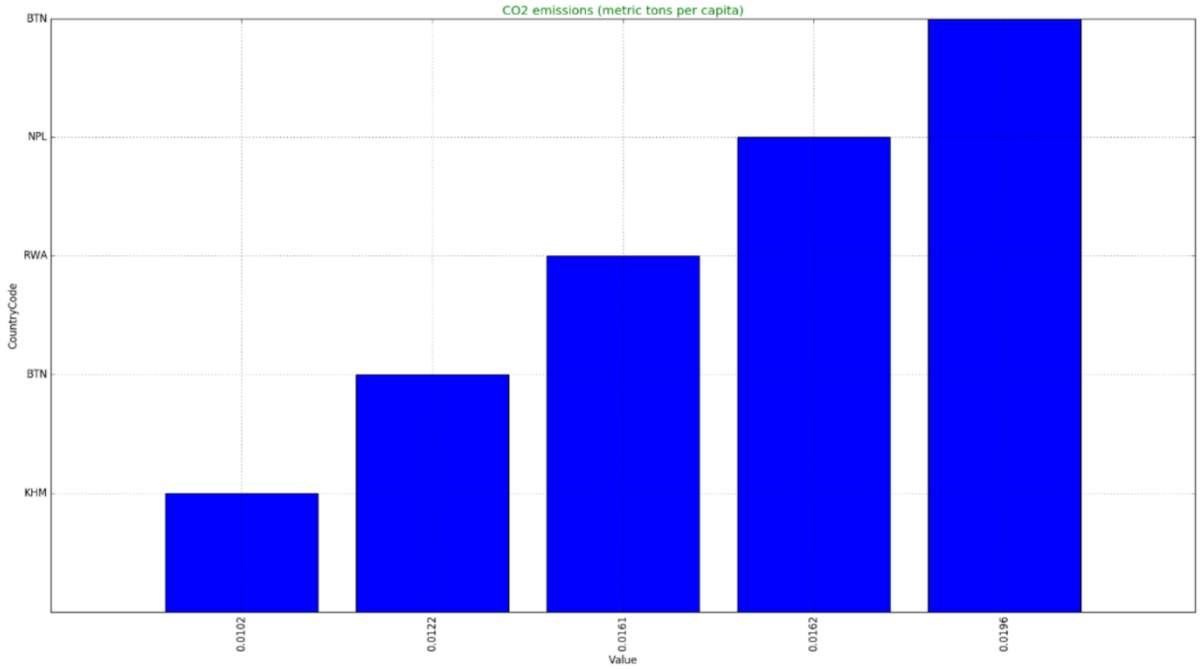
Here are comparison of CO2 emissions between Brazil and India



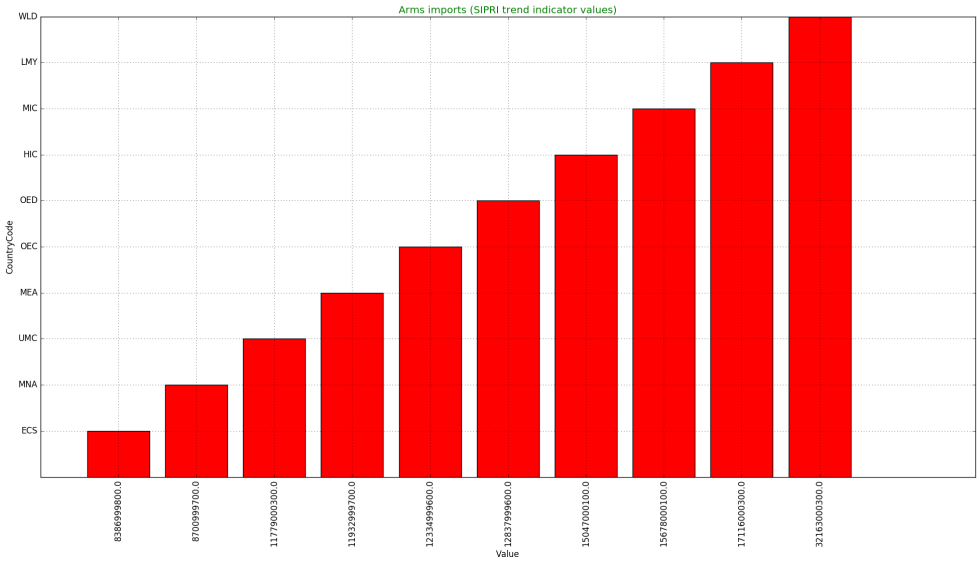
Visualisation of results observed

Below is the graphical representation of each of the reports

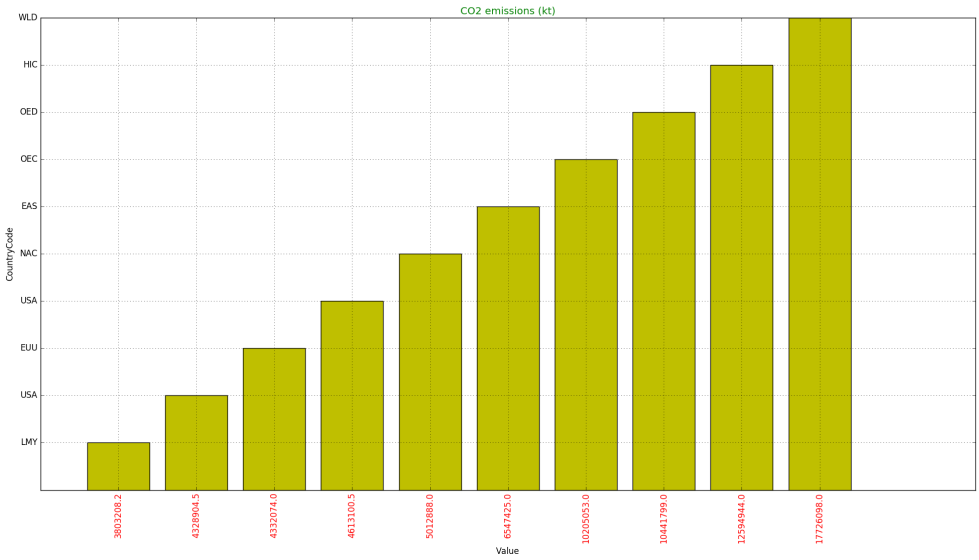
a. Top 10 countries which are exporting Arms



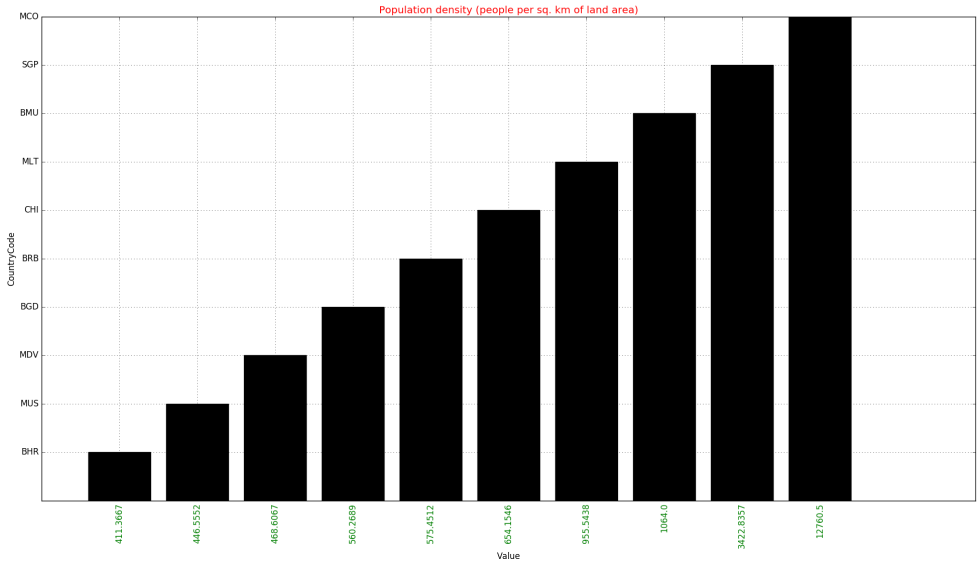
b. Top 10 countries which are importing Arms



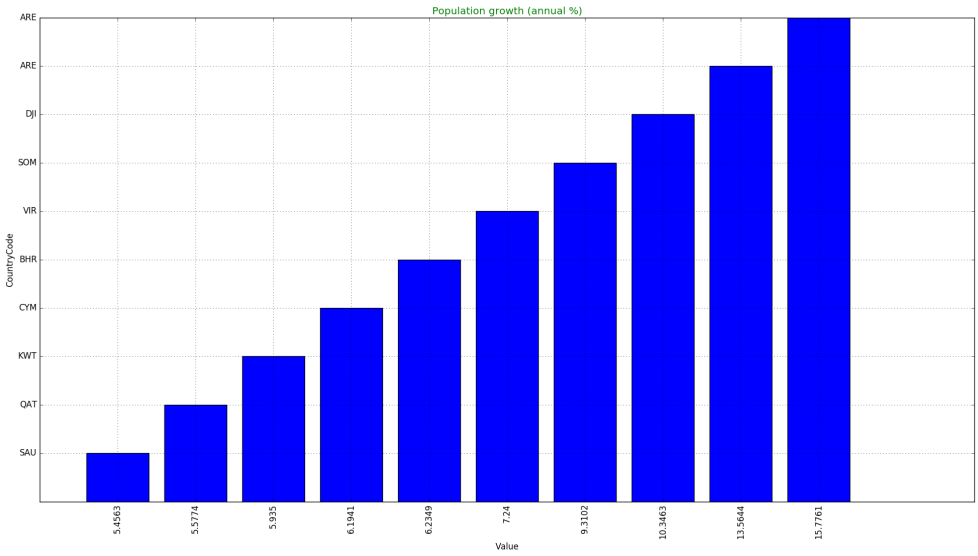
c. Top 10 countries which are emitting CO2.



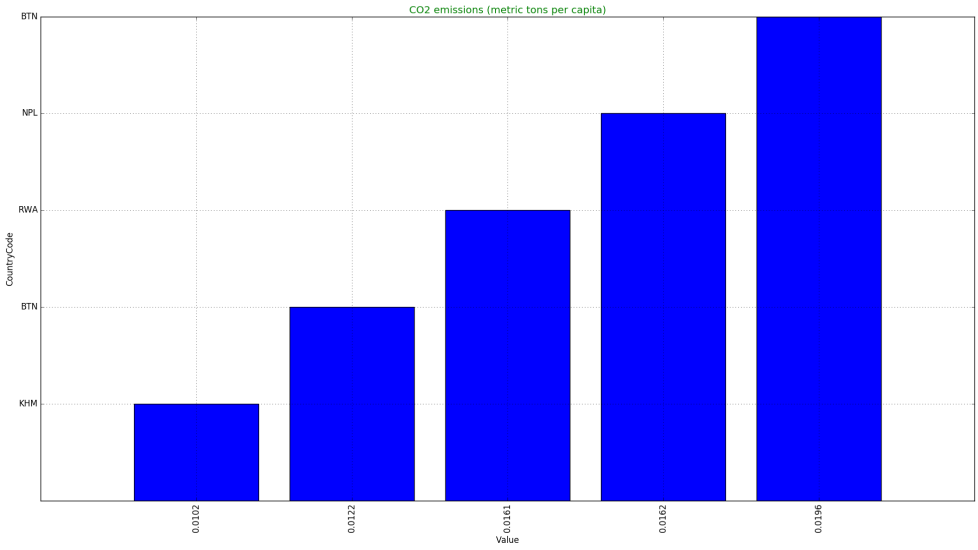
d. Top 10 countries which are having highest Population density.



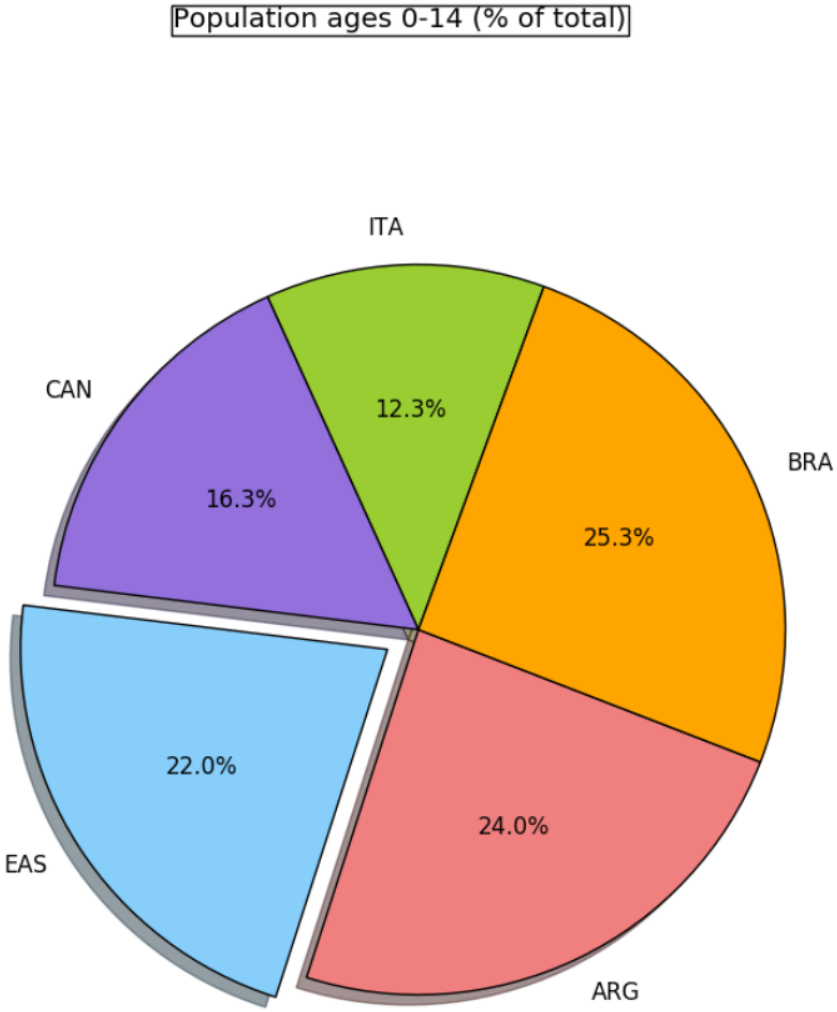
e. Top 10 countries which are having population growth.



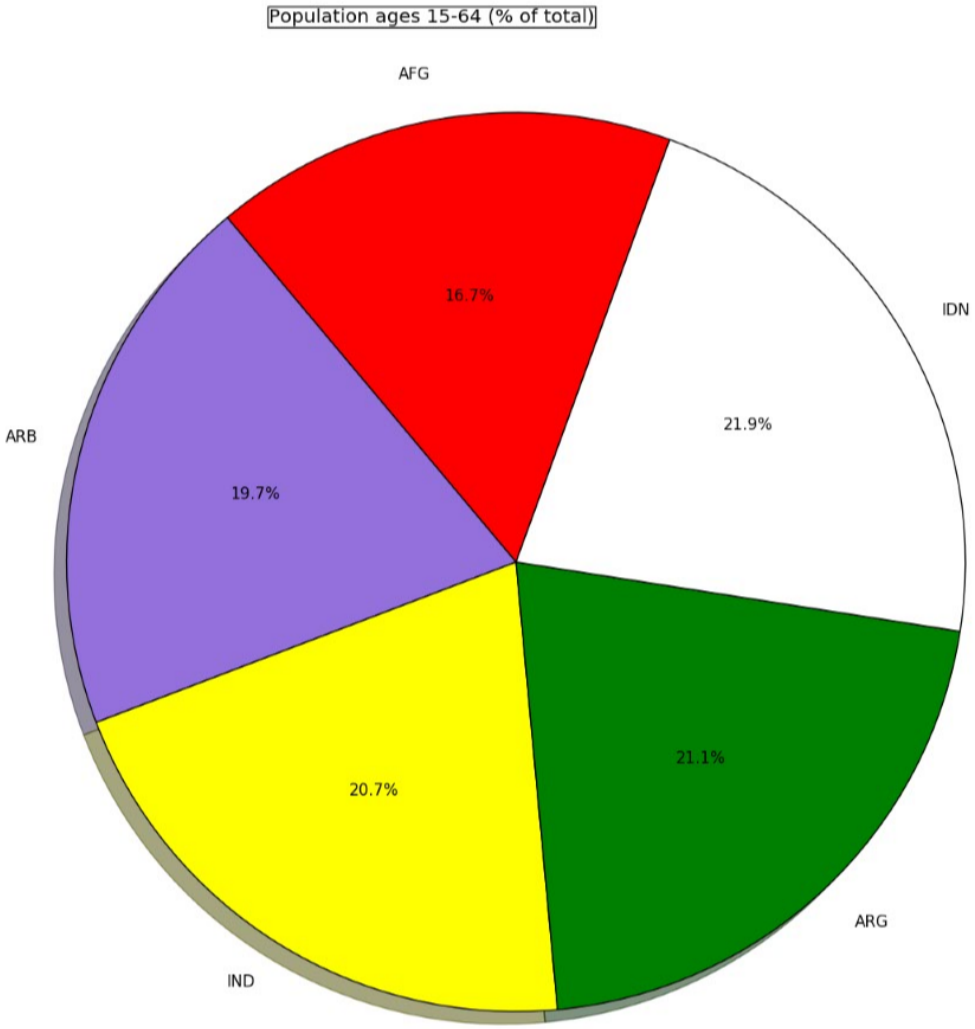
f. Least 5 countries which are emitting low CO2.



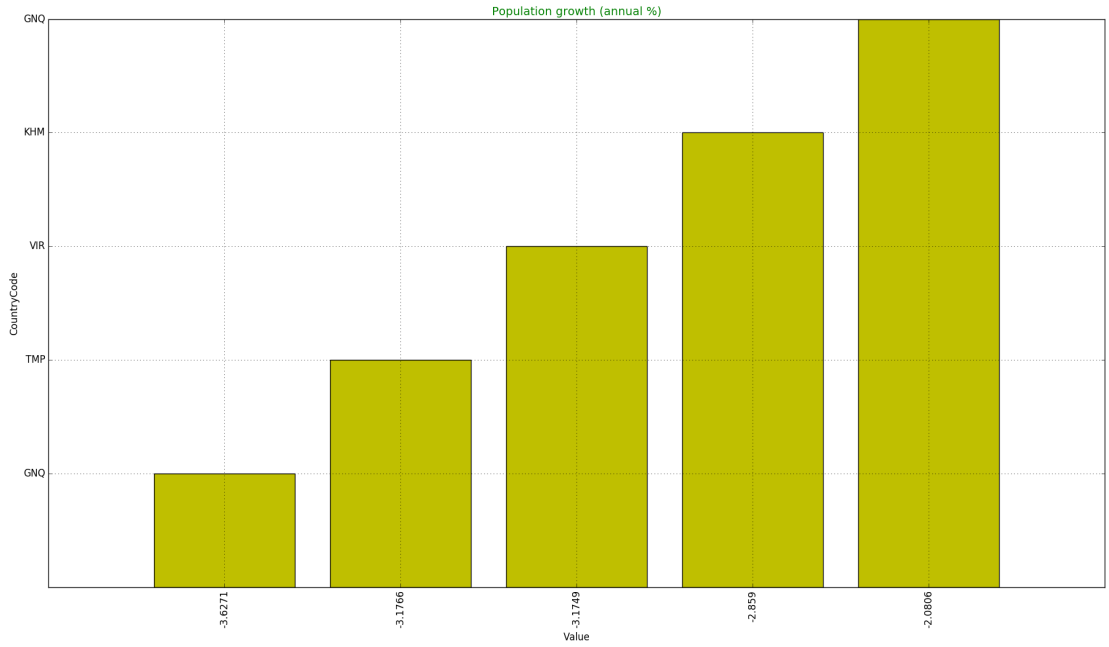
g. Least 5 countries with low Population between 0-14 years (% of total).


Population ages 0-14 (% of total)
ITA
12.3%
BRA
25.3%
CAN
16.3%
EAS
22.0%
ARG
24.0%

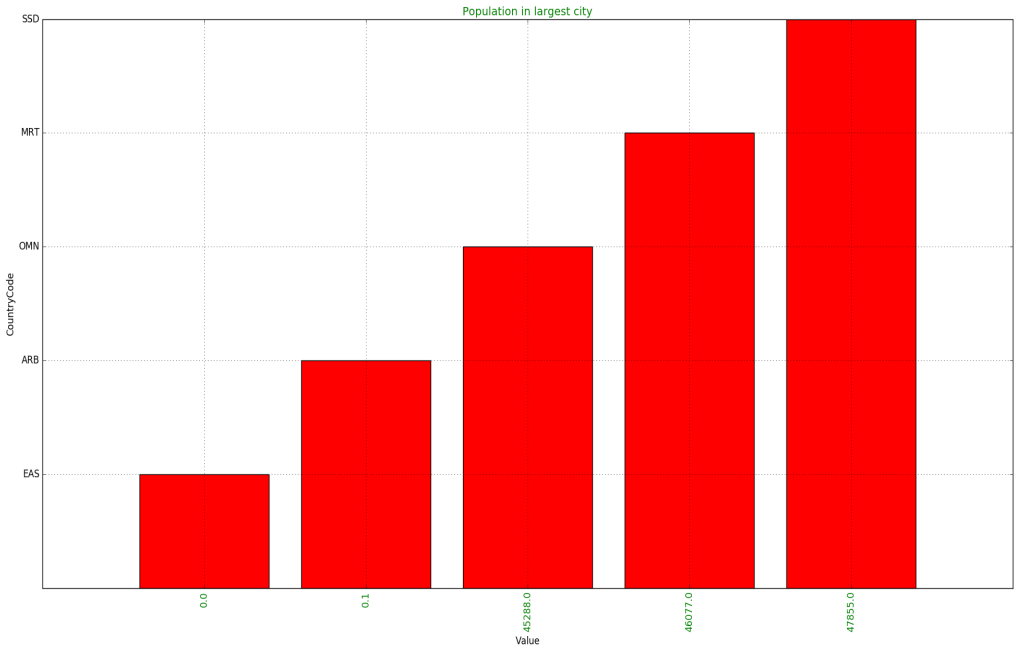
h. Least 5 countries with low Population between 15-64 years (% of total)



i. Least 5 countries with low Population growth



j. Least 5 countries which are having largest population



Program and Output files:

Please find below are the program and output files which are used in this project.

Conclusion:

It is important to understand the current statistics of different sectors of developing countries and their positioncomparable worldwide to make the proper plan and decision. In our project through top-n analysis we have derived the top and least 5 position of the countries for selected sectors which would be the clear indication of steps takento improve the quality of people life.