INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT

(Project Semester August-December 2020)

Analysis on Gross Domestic Product data of various countries

Submitted by

Arjun Pandey

11810731

KM078

B Tech. (Computer Science and Engineering)

Course Code: **INT217**

Under the Guidance of

Miss. Savleen Kaur UID: 18306

Discipline of CSE/IT

School of Computer Science & Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Arjun Pandey** bearing Registration no. **11810731** has completed **INT217** project titled, "***Analysis on Gross Domestic Product data for various countries***" under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 18th December 2020

DECLARATION

I, **Arjun Pandey** student of **BTech CSE**, under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 18th December 2020

Registration No.11810731 Arjun Pandey

Acknowledgement

In preparation of my project, I had to take the help and guidance of some respected persons, who deserve my deepest gratitude. As the completion of this assignment gave me much pleasure, I would like to show my gratitude **Miss. Savleen Kaur**, Course Instructor, in Lovely Professional University for giving me a good guideline for project throughout numerous consultations. I would also like to expand my gratitude to all those who have directly and indirectly guided me in writing this report.

Many people, especially my classmates have made valuable comment suggestions on my project which gave me an inspiration to improve the quality of the assignment.

Table of Content

Introduction

Before taking about anything, at first I will answer all three common question of **what, why** and **how** regarding data analysis.

What is Data Analysis?

Data analysis is defined as a process of cleaning, transforming, and modelling data to discover useful information for business decision-making. Whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analysing our past or future and making decisions based on it. For that, we gather memories of our past or dreams of our future. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

Why Data Analysis?

To grow your business even to grow in your life, sometimes all you need to do is Analysis!

If your business is not growing, then you have to look back and acknowledge your mistakes and make a plan again without repeating those mistakes. And even if your business is growing, then you have to look forward to making the business to grow more. All you need to do is analyse your business data and business processes.

How to do Data Analysis?

Using- Data Analysis Tools

Data analysis tools make it easier for users to process and manipulate data, analyse the relationships and correlations between data sets, and it also helps to identify patterns and trends for interpretation. Tools for doing analyse including MS Excel, Tableau, Python Programming, R Programming etc.

Here in this project, I will be taking a dataset of GDP worldwide with respect to the year ranging between 1970 to 2019 and draw some conclusion about which countries are lying where in the world in terms of GDP growth.

SCOPE OF THE ANALYSIS

- To analyze GDP growth of country and obtain a trend line.
- To analyze that when the world has gone into recessions.
- To analyze and know total GDP per category.
- To analyze and know the total GDP produced by each country.
- To make a visualization in Tableau using world map and placing details respectively.

Source of Dataset

Data has been downloaded from **worldbank.com**

All country GDP data from 1970 to 2019

Context

The dataset will be valuable to those who seek to understand the dynamics of gross domestic product of various countries.

Content

The dataset specifically focuses on the gross domestic production of countries in the course of a time. This dataset consists of many countries GDP data.

The data contains 55 columns and 214 rows making a total of 11,770 number of cells.

The data set includes information about:
The country and their respective GDP generation from 1970 to 2019.
data.xlsx

- Country Code: This code is given to every country and is unique for every country.
- Income Group: This is the categorization of the countries under tags such as Lower middle income, Low income, High income and Upper middle income class.
- Country: This is the name of the countries.
- Sum: The total sum of GDP produced by each country in whole timespan.
- Average: The average of GDP produced by respective country.
- Years: This ranges from 1970 to 2019.
- Sum yearly: Total GDP produced year wise including all country.

ETL Process

What is ETL?

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.

It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

Why ETL?

There are many reasons for adopting ETL in the organization:

- It helps companies to analyse their business data for taking critical business decisions.
- Transactional databases cannot answer complex business questions that can be answered by ETL.
- A Data Warehouse provides a common data repository
- ETL provides a method of moving the data from various sources into a data warehouse.
- As data sources change, the Data Warehouse will automatically update.
- Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.
- Allow verification of data transformation, aggregation and calculations rules.
- ETL process allows sample data comparison between the source and the target system.
- ETL process can perform complex transformations and requires the extra area to store the data.
- ETL helps to Migrate data into a Data Warehouse. Convert to the various formats and types to adhere to one consistent system.
- ETL is a predefined process for accessing and manipulating source data into the target database.
- ETL offers deep historical context for the business.

- It helps to improve productivity because it codifies and reuses without a need for technical skills.

ETL is a 3-step process

Step 1) Extraction

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Data warehouse needs to integrate systems that have different DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, Point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others.

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

Some validations are done during Extraction:

- Reconcile records with the source data
- Make sure that no spam/unwanted data loaded
- Data type check
- Remove all types of duplicate/fragmented data
- Check whether all the keys are in place or not

Step 2) Transformation

Data extracted from source server is raw and not usable in its original form. Therefore, it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as direct move or pass through data.

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

Following are Data Integrity Problems:

- Different spelling of the same person like Jon, John, etc.
- There are multiple ways to denote company name like Google, Google Inc.
- Use of different names like Cleaveland, Cleveland.
- There may be a case that different account numbers are generated by various applications for the same customer.
- In some data required files remains blank
- Invalid product collected at POS as manual entry can lead to mistakes.

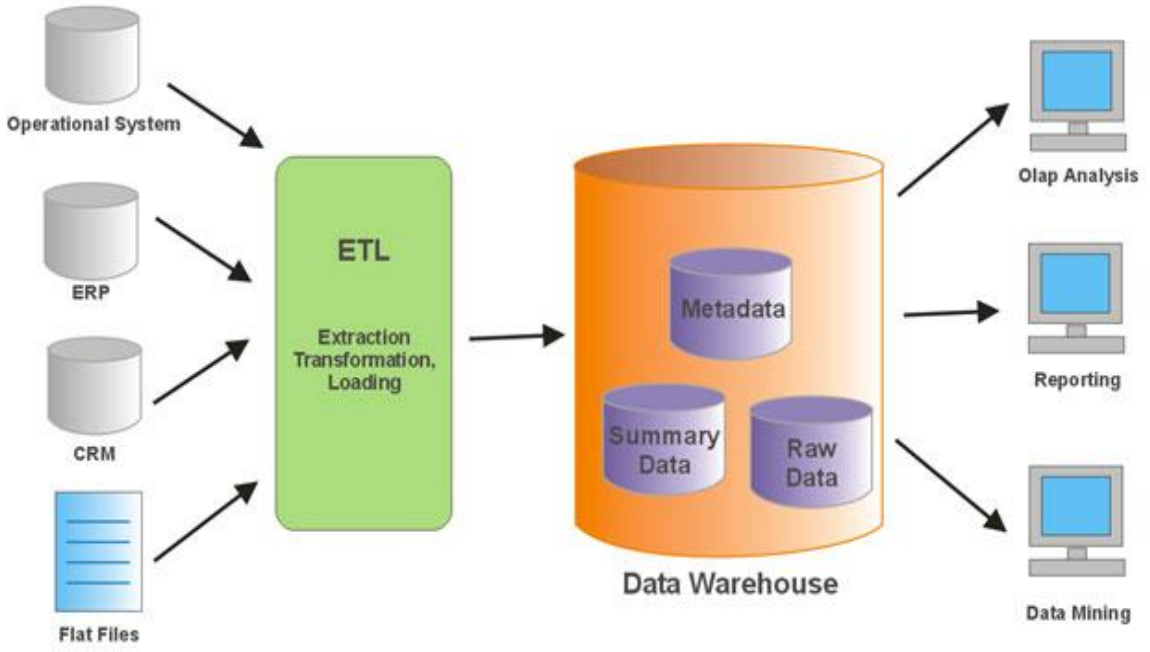
Validations are done during this stage

- Filtering – Select only certain columns to load
- Using rules and lookup tables for Data standardization
- Character Set Conversion and encoding handling
- Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
- Data threshold validation check. For example, age cannot be more than two digits.
- Data flow validation from the staging area to the intermediate tables.
- Required fields should not be left blank.
- Cleaning (for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)
- Split a column into multiples and merging multiple columns into a single column.
- Transposing rows and columns,
- Use lookups to merge data
- Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically rejects the row from processing)

Step 3) Loading

Loading data into the target data warehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.



Load verification

- Ensure that the key field data is neither missing nor null.
- Test modelling views based on the target tables.
- Check that combined values and calculated measures.
- Data checks in dimension table as well as history table.
- Check the BI reports on the loaded fact and dimension table.

ANALYSIS ON DATASET

ANALYSIS 1:

INTRODUCTION:

To analyse GDP growth of country and obtain a trend line..

GENERAL DESCRIPTION:

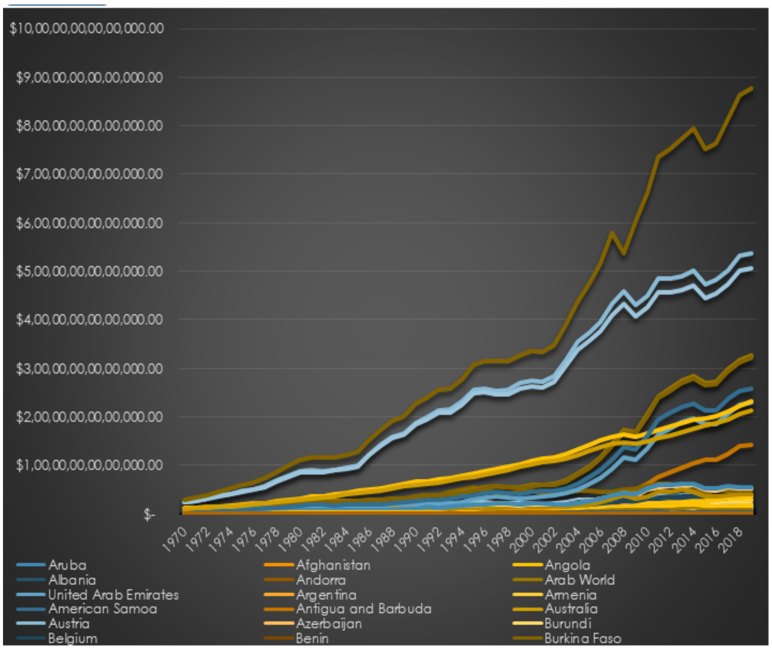
This analysis is to help us to understand the trend line according to which each country is growing it's GDP as the time is passing by.

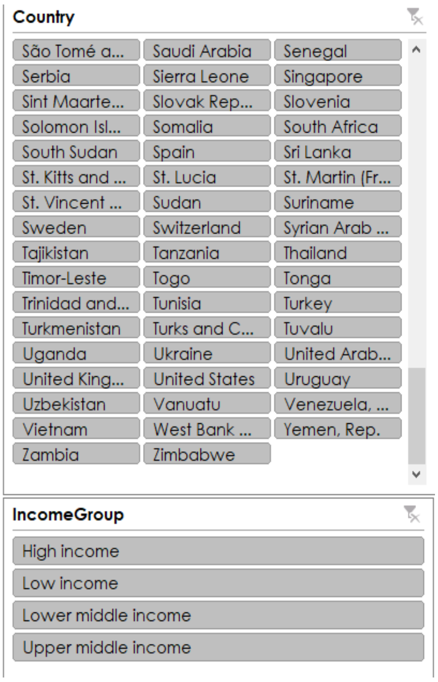
SPECIFIC REQUIREMENTS:

The main requirement is to have the list of the countries and the data for each country yearly GDP.

The main dataset is used to make this visualization and some slicers through which we can get a personalized experience and get trend line as per country selected. Countries is placed in row label and GDP is placed in column label.

VISUALISATION:





ANALYSIS 2:

INTRODUCTION:

This visualization analyses that when the world has gone into recession.

GENERAL DESCRIPTION:

This analysis is to help the analyst to understand when there were recessions in the global economy and have they even recovered or not.

SPECIFIC REQUIREMENTS:

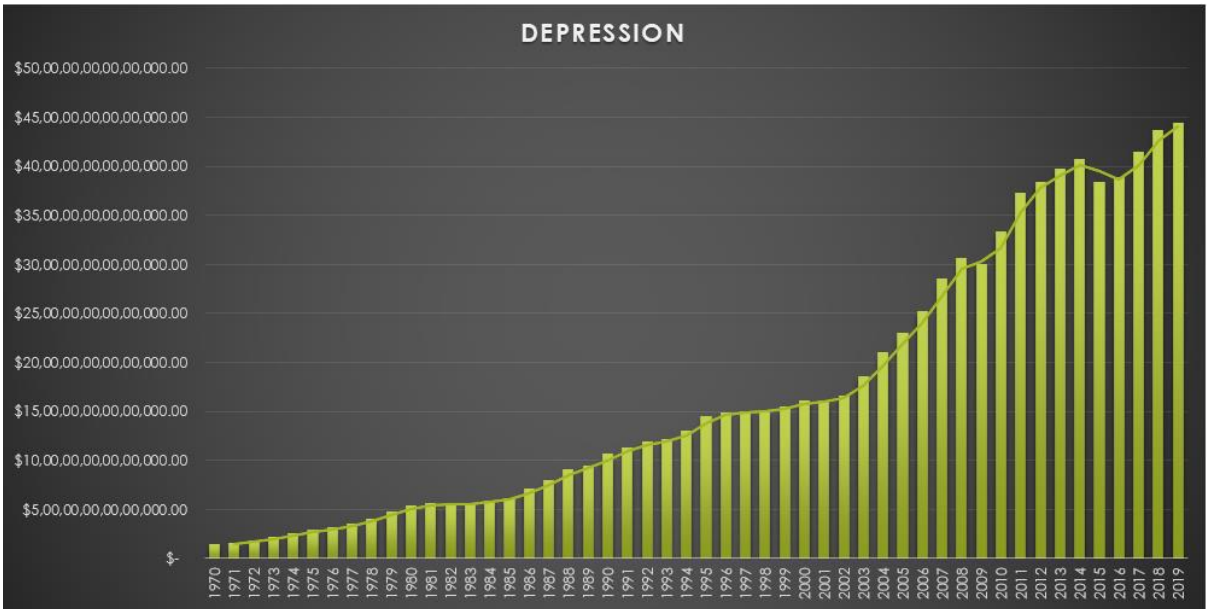
The main requirement is to have the sum of GDP year-wise.

This visualization is being made on main dataset and here the years are listed in X-axis and the GDP in Y-axis upon which we are putting a regression line calculated on moving averages through which we can predict what should have been the GDP.

ANALYSIS RESULTS:

Using the analysis through the visualization we came to a conclusion that there were two time the world GDP went into recession i.e. 2008-2009 and 2015-2016 that can be clearly inferred from the chart below.

VISUALISATION:



ANALYSIS 3:

INTRODUCTION:

To analyse and know total GDP per category and their contribution to the net GDP of the world.

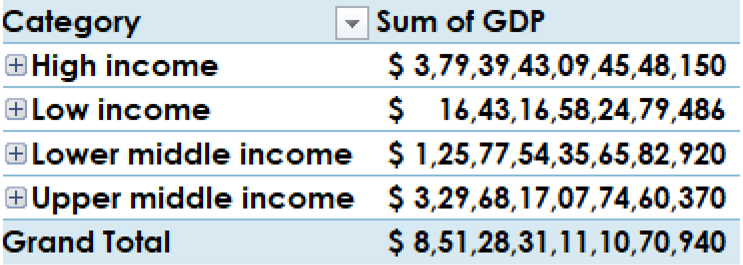
GENERAL DESCRIPTION:

This analysis will help the analyst to figure out which category of country is contributing what percentage and what amount to the world's economy and according to that take preventive measures to improve it.

SPECIFIC REQUIREMENTS:

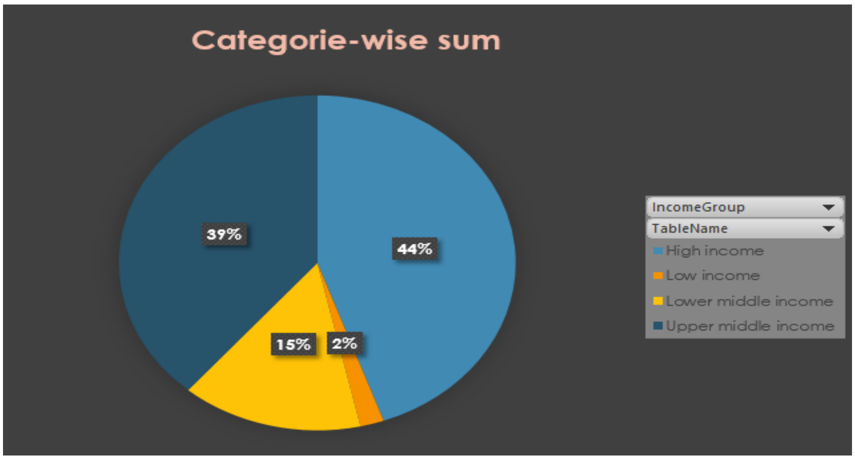
The main requirement is to have the list of the countries inside their particular category. Pivot table is used to calculate the percentage and amount of GDP contributed by each category. Countries is placed in row label and Banking Crisis in column label.

ANALYSIS RESULTS:



Category	Sum of GDP
⊞High income	$ 3,79,39,43,09,45,48,150
⊞Low income	$ 16,43,16,58,24,79,486
⊞Lower middle income	$ 1,25,77,54,35,65,82,920
⊞Upper middle income	$ 3,29,68,17,07,74,60,370
Grand Total	$ 8,51,28,31,11,10,70,940

VISUALISATION:



ANALYSIS 4:

INTRODUCTION:

To analyse and know the total GDP produced by each country.

GENERAL DESCRIPTION:

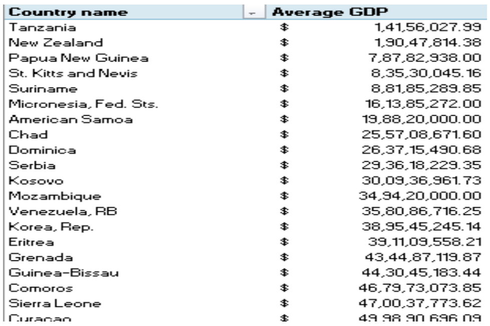
This analysis will help Analysts to understand the data and find which country is holding the position of highest GDP producing country as well as which country is having the position of lowest GDP producing country.

SPECIFIC REQUIREMENTS:

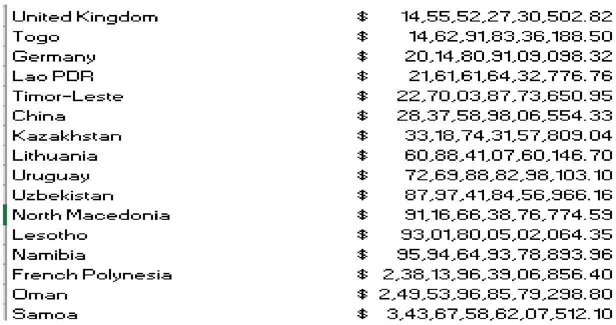
The main requirement is to have the list of the countries and the average of their GDP.

Pivot table is used to calculate the GDP on different countries.

ANALYSIS RESULTS:

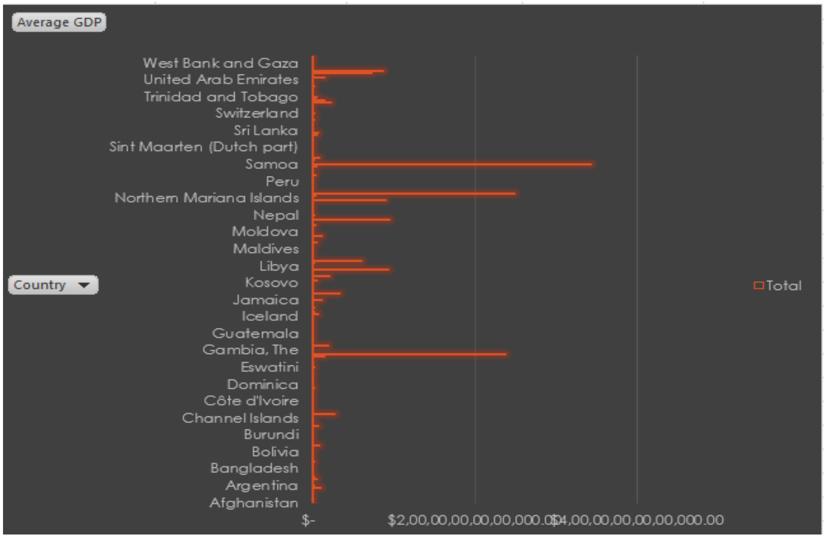


Country name	Average GDP	
Tanzania	$	1,41,56,027.99
New Zealand	$	1,90,47,814.38
Papua New Guinea	$	7,87,82,938.00
St. Kitts and Nevis	$	8,35,30,045.16
Suriname	$	8,81,85,289.85
Micronesia, Fed. Sts.	$	16,13,85,272.00
American Samoa	$	19,88,20,000.00
Chad	$	25,57,08,671.60
Dominica	$	26,37,15,490.68
Serbia	$	29,36,18,229.35
Kosovo	$	30,09,36,961.73
Mozambique	$	34,94,20,000.00
Venezuela, RB	$	35,80,86,716.25
Korea, Rep.	$	38,95,45,245.14
Eritrea	$	39,11,09,558.21
Grenada	$	43,44,87,119.87
Guinea-Bissau	$	44,30,45,183.44
Comoros	$	46,79,73,073.85
Sierra Leone	$	47,00,37,773.62
Curacao	$	49,98,90,696.09



United Kingdom	$	14,55,52,27,30,502.82
Togo	$	14,62,91,83,36,188.50
Germany	$	20,14,80,91,09,098.32
Lao PDR	$	21,61,61,64,32,776.76
Timor-Leste	$	22,70,03,87,73,650.95
China	$	28,37,58,98,06,554.33
Kazakhstan	$	33,18,74,31,57,809.04
Lithuania	$	60,88,41,07,60,146.70
Uruguay	$	72,69,88,82,98,103.10
Uzbekistan	$	87,97,41,84,56,966.16
North Macedonia	$	91,16,66,38,76,774.59
Lesotho	$	93,01,80,05,02,064.35
Namibia	$	95,94,64,93,78,893.96
French Polynesia	$	2,38,13,96,39,06,856.40
Oman	$	2,49,53,96,85,79,298.80
Samoa	$	3,43,67,58,62,07,512.10

VISUALISATION:



ANALYSIS 5:

INTRODUCTION:

To make a visualization in Tableau using world map and placing details respectively.

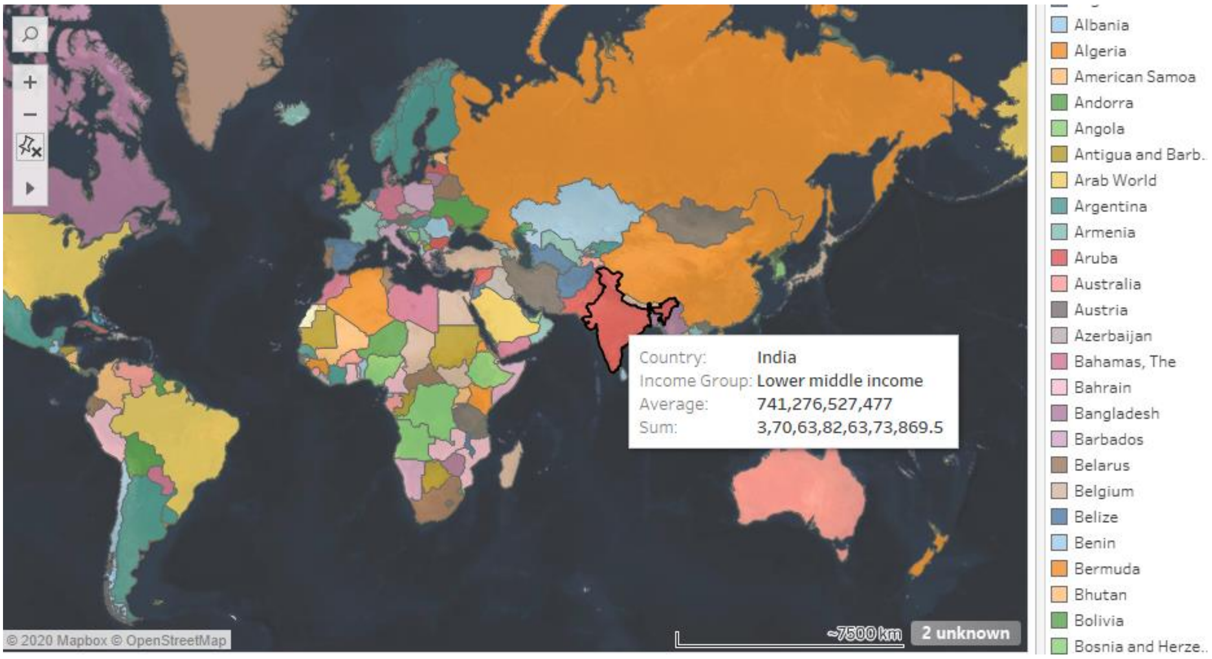
GENERAL DESCRIPTION:

This will help the analysts to easily navigate through each and every country on the world map and know its related data and infer accordingly.

ANALYSIS RESULTS:

We come to know that the GDP produced by each and every country in an easy manner.

VISUALISATION:



Country: India
Income Group: Lower middle income
Average: 741,276,527,477
Sum: 3,70,63,82,63,73,869.5

Future Scope

Data analytics is a process through which data is cleaned, analysed and modelled using tools. This data is then used to derive insights. The insights are then used for business-related decision-making purposes. There are many techniques that data analysts use in different fields of work. In the world of business, Data analytics is used for making strategies to get the desired business results. Today, data analytics has become a big career option in India. As a result, big data analytics courses are in huge demand.

Businesses have realised the importance of utilising big data analytics to maximize their profits. They know that it is vital for their growth and for the future health of their business. Today, major business decisions are taken by utilising the insights derived from data related to the organization or industry-related data. As competition increases and customers are flooded with choices, it has become important to move faster in the market and that too with accuracy.

Data analytics provides both speed and accuracy to business decisions. It provides accuracy as it is based on statistical models and hi-tech tools that help fine-tuning and analysing the data. This field also provides answers to present business problems as well as give a view of future trends. It is preparing the companies to make products for the future and aspire to connect with the customers of tomorrow.

As data analytics also allows to improve business process and maximise conversion rates, it helps the organizations in cutting unnecessary costs and reduce the cost of running the company. With all its obvious benefits, it is quite natural to say that data analytics is going to become important in a big economy like India.

India is a popular destination for a lot of companies who outsource their work to other countries. This is due to the lower cost of operations and manpower in India. This is further aided by the skilled and English-speaking youth of India. Data analytics is one such field where outsourced opportunities are available in India. As a country teeming with young people and tremendous outsourced work coming in, the scope for this sector is big in India.

Today, as advancements in the field of data analytics are being made, the process is getting automated. Machines are analysing big chunks of data in an automated process. With more and smarter machines entering our daily lives, more and more data is getting created every hour. All this data can be used and analysed for understanding customer behaviour or predicting future trends. With the help of machines, data analysts are finding it possible to make sense of the data in a quicker and easier way.

This is true of India as well. Data around us is growing at a very fast rate. This is because of the changes that the country is going through. The smartphones and data plans are getting cheaper, data speeds are getting faster, and social media is becoming a trendy way of connecting with friends or voicing one's opinions. All these changes are generating a lot of data around us and organizations realise that all this data can be cleaned and analysed to find useful information. For example, Google uses the data it gets from our smartphones to understand the movement of traffic on our streets. The information helps in providing its user's information about the distance and time taken to reach their destination in real-time through the Google Maps app.

With newer technologies on the horizon, words like Block chain, Internet of Things, machine learning, Artificial Intelligence etc. have been the most popular lexicons among business corridors. The most interesting thing about all the modern technology is that they are all based on data.

Because of the bright future of data analytics, many professionals and students are interested in a career in data analytics. Any person who likes to work on numbers has logical thinking, can understand figures and can turn them into actionable insights, has a good future in this field. Proper training of the tools of data analytics would be required to begin with. Since it is a course that requires effort to learn and get certified, there is always a dearth of qualified professionals. Being a relatively new field also, the demand for such professionals is more than the current supply. Higher demand also means higher salaries.

Data analytics is the differentiator that provides companies with a competitive edge over others. It is a fast-growing branch of study which has a bright future in India. Organizations have realised their importance and investing in data analytics tools and technologies. Professionals and students are keen for a career in data analytics owing to the career opportunities they might get. We can be sure that data analytics has a good future in India for years to come.

Bibliography

1) Blog: Scope and Future of Data Analytics in India

https://talentedge.com/blog/scope-future-data-analytics-india/

2) Blog: What is Data Analysis? Types, Process, Methods, Techniques

https://www.guru99.com/what-is-data-analysis.html#1

3) Blog: ETL (Extract, Transform, and Load) Process

https://www.guru99.com/etl-extract-load-process.html

4) Blog: A Guide to Basic Data Analysis

https://www.geckoboard.com/learn/data-literacy/basic-data-analysis-guide/

5) Blog: The 4 Important Things About Analysing Data

https://blog.treasuredata.com/blog/2015/04/03/4-important-things-about-analyzing-data-understand-the-purpose-and-results/

6) Blog: Data Analysis

https://en.wikipedia.org/wiki/Data_analysis

7) Concept of Excel Pivot

https://www.excel-easy.com/data-analysis/pivot-tables.html

8) Concept of Multi Level Pivot Table

https://www.excel-easy.com/examples/multi-level-pivot-table.html

9) Concept of Slicers

https://www.excel-easy.com/examples/slicers.html

10) Concept of Pivot Charts

https://www.excel-easy.com/examples/pivot-chart.html