DATASET ANALYSIS FOR FINDING BEST PLACES FOR LIVING

Ms. Arranya Balaram[1], Dr.habil.sc.ing. Janis Grundspenkis[2]

[1]Student, M.S Engineering Science in Computer Systems,Riga Technical University,1 Kalku Street, LV-1658 (Riga)(Latvia)

[2]Professor and Head of the Department of Artificial Intelligence and Systems Engineering and Dean of the Faculty of Computer Science and Computer Systems,Riga Technical University,1 Kalku Street, LV-1658 (Riga)(Latvia)

---***---

Abstract – *The master thesis utilizing the big data analytics process to use them together with Tableau visualization tool for predicting the possible results in the form of best livable countries fits for survival. The thesis contains workflow of the project and its implementation process. The master thesis describes visualization results of each dataset connected into Tableau public software based on the concepts of quality of life and standard of life statistical data. Datasets information using for the thesis are selected and collected by the author of the thesis from the data world bank web resource. Analysis of the datasets according to the public preference for finding best livable countries are analyzed in the form of visualization report. The results of countries which are achieved from the master thesis report can be bit similar to some list of most livable countries/cities in the world which are published every year by Economist Intelligence Unit (EIU) under Global Livability Ranking countries.*

***Key Words*: Big data analytics, Visualization tool, Tableau, Statistical datasets.**

1. INTRODUCTION

Nowadays, the usage of big data increased constantly by the users in the form of utilizing various web resources for various purposes. This thesis presents the idea of using visualization tool for implementing statistical datasets related to the quality and standard of living which are directly collected from world bank website. The project is based on big data analysis, the public can find out the best fittest countries for their survival with the help of Tableau public visualization tool. It is open source, user-friendly, free of cost for installation. No programming knowledge is required. So even public can get best visualization reports. This tool allows the user to connect and visualize datasets in a meaningful way. The advantage of using this tool highlights the visualization pattern type for analysing report. The analysis results of all reports can also be combined in the form of story creation. This representation helps users to obtain conclusion by comparing results with each report in a convenient manner.

The goal of the master thesis is to produce best results about the suitable countries for living. This thesis work is focused on people who are searching best countries to live. Public perception is based on choosing countries with good medical facilities, education facilities, affordable living cost, safety and security, government support etc., To achieve this goal, the following tasks are defined.

- defining the specification using for the analysis of the project;
- defining the requirement of search according to the public point of view questions;
- collecting required datasets from the data source for implementation into Tableau tool;
- filtering and arranging the data information for each requirement specification;
- exploring the datasets in the form of visualization representation which provides detailed statistical information of each country in a meaningful way.

2. BIG DATA ANALYTICS OVERVIEW

The challenge of this era has to deal with large volume of data increased rapidly to unimaginable levels. Concurrently, in the past decade, data storage price has been reduced systematically. Basically, terabytes to exabytes/petabytes of data-driven by private companies for business and also by research institutions utilizing data for finding the details about their users and also for other purposes like interactions on social media. On the other hand, real-time sensors can be used to detect from physical devices such as mobile phones, automobiles etc., To make sense of this sea of data brings big data analytics into picture.

Data Layer Fig-1 Relational database management system (RDBMS) consists of structured, unstructured and semi-structured data information are retrieved and analyzed with the implementation of the visualization tool. For example, NoSQL database like MongoDB and Cassandra and also the stream of data taken from web resources such as IoT, social media are used to store unstructured data. Some software tools such as Flume, Kafka, Pig, Scala, Spark, Storm, Tajo, HBase, Impala, Hcatalog, Hive, MapReduce, Teradata, Sqoop, PowerBL, Hadoop, Cassandra, SAS support this layer.

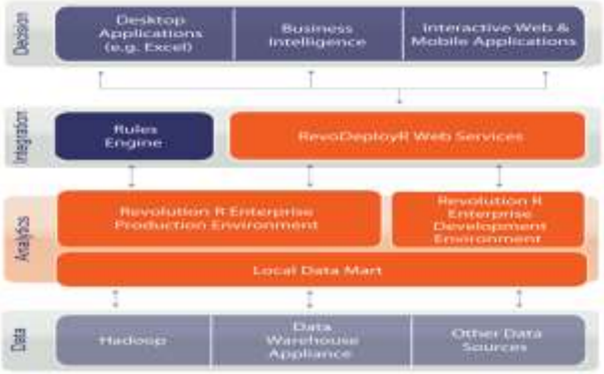

Decision
Desktop Applications (e.g. Excel)
Business Intelligence
Interactive Web & Mobile Applications
Integration
Rules Engine
RevoDeployR Web Services
Analytics
Revolution R Enterprise Production Environment
Revolution R Enterprise Development Environment
Local Data Mart
Data
Hadoop
Data Warehouse Appliance
Other Data Sources

Fig -1: Big data analytics architecture

Analytics Layer improves the performance of the visualization tool. It helps to implement both static and dynamic data analytics in the environment to develop models, to modify local data and to deploy real-time values.

Integration Layer contains analytics layer together integrates an application program interface (API) with an end-user application

Decision Layer decides the end products with the help of end-user software such as web application, desktop app and business intelligence apps which interacts with the system.

3. JUSTIFICATION OF EXISTING AND SELECTED VISUALIZATION TOOLS

Data visualization represents the extraction of raw data collected from various data sources using visualization tools. Structure of the text and interrelationships can be easily visualized by applying some suitable visualization patterns such as graphs, charts, plots, histogram etc., to the datasets implemented into the worksheet/dashboards. The main advantage of data visualization in as follows:

3.1 Benefits of visual representation of data

Enhance better decisions Applications can be able to analyze and visualize the condition of their own selection details which includes user policies, rating-standards, sharing tasks with other web users to improve better by making appropriate decisions and apply the changes in their activities.

Enhance return on investment(ROI) Applications can be able to analyze and visualize the advantages and disadvantages of their activities. It is very easy to predict miscalculations and it can get resolved to enhance the return on investment.

Sharing information Applications can be able to view their preceding and existing activities. By taking the user opinions and decision makers analysis results into consideration for the report and applying those gathered analysis information changes in the report for their future use. They can also share their information to keep away from misinterpretation.

Saving time the great benefit of visualization is to solve the problem in a limited period of time. In place of the experimental approach, this approach gives an instant solution to related data and recover the insights in visual data.

3.2 Analysis of existing tools in big data

Discussion of some of the popular tools existed in the data analysis field used by wide range of users/data scientists are as follows:

Table -1: Representation of attributes

	Suitable working platform	**Required skill**	**Data access limits**	**Utilization of data sources**	**Learning tutorials**
Google fusion tables	Web browser	Beginner	Limited data	Yes	Yes
Microsoft PowerBL	Windows and web browser	Beginner	Handles complex	Yes	Yes
Tableau public	Windows and MacOS	Advance beginner to intermediate	Handles complex	Yes	Yes
Google data studio	Web browser	Beginner	Limited data	Yes	Yes
Plotly	Web browser	Beginner for windows and experts for online service	Handles complex	Yes	Yes
Qlik sense desktop	Windows (64-bit)	Advance beginner to intermediate	Handles complex	Yes	Yes
VIDI	Web browser	Beginner	Limited data	Yes	Yes
Zoho reports	Web browser	Advance beginner	Limited data	Yes	Yes
IBM word cloud generator	Windows, MacOS, Linux	Advance beginner	Handles complex	Yes	Yes
Time flow	Windows, MacOS,	Beginner	Handles complex	Yes	Yes
Open street map	Web browser	Advance beginner to intermediate	Handles complex	Yes	Yes

3.3 Free of error in data visualization

During visualization of data, when dealing with complication data may discovers some trouble when presenting a report. The below discussion helps to understand how to tackle and manage the huge volume of data.

Publishing entire data. Fault may occur when fetching the data in visualization tool. It always represents precise data to the users/data analysts. Usually, data information is organized in a well-defined structure with meaningful information. But some of the data retrieved from data source are organized in a bad structure. When viewing that information in the dashboard it seems to be nuclear and it is very difficult to understand the meaning of visual data.

Misconception in view. Data visualization helps the users to take better decisions. This cumbersome visual data leads user/data analyst to take a wrong decision.

Deficient plan. For presenting data, user/data analysts need to ensure that they have whether chosen suitable dashboards or not. Due to lack of idea or prior plan about the work users are facing trouble with the quality of the view.

3.4 Design flow of file types and data types

Fig – 2: illustrates the design flow of the Tableau tool. The explanation of each flow and how it works as discussed below:

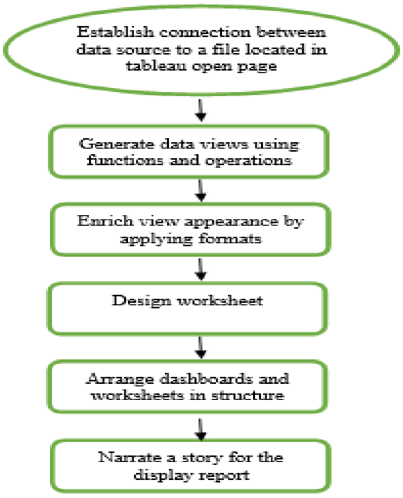

Establish connection between data source to a file located in tableau open page
Generate data views using functions and operations
Enrich view appearance by applying formats
Design worksheet
Arrange dashboards and worksheets in structure
Narrate a story for the display report

Fig -2: Visualization report

In Tableau visualization tool, initially, the user needs to connect existing datasets that are available in their system or which is already collected from different data sources. These files have to connect first with the appropriate file such as Excel text file, access. JSON file, PDF, spatial or statistical files or if the user project is server based then connect to Google sheets, OData or web data connector. Tableau public edition has been integrated with data connectors which helps to gather and validate the wide range of standard database such as relational, big data, salesforce, Google analytics etc., After getting into work page user can place sheets by double-clicking onto to drag sheet area. The user can sort or align data views according to their convenience. When opening the worksheet area user can see set of panels which hold data shelves, visualization patterns etc., User can perform various functions and operations like filtering, extracting data to enrich the appearance of the worksheet using dimensions and measures field. Users can work with multiple worksheets and also with dashboards by creating new worksheet/dashboard. In this area, the user can design the worksheet with the help of features built-in under worksheet/dashboard command menus. Then arrange the worksheets or dashboards for analyzing data. The user can also narrate a story with the collection of all worksheets and dashboards

4. COLLECTION OF APPROPRAITE DATASETS FOR FINDING BEST PLACES FOR SURVIVAL

The data survey is based on the superior quality of developing countries, fittest countries place for the human to live, the measure of data based on the ranking values. The aim of the research is based on analyzing various datasets on the basis of quality, ranking and living standard from the data world bank web resource.

4.1 Factors of quality and standard of living

The *quality of living* refers to the individual or social welfare. The factors which include in quality of life are public safety, education facilities, transport facilities, economical resources, political supports, safe environment, health, income, infrastructure facilities.

The *standard of living* refers to the particular countries which public can be able to know the details how countries providing good pay to the employees, comfortability in lifestyles, commonly services. The standard of living is similar to the quality of life. It comes under such factors which includes value of national income, employment statistics, production of goods and services, population density, literacy dates, life expectancy, education level, health care accessibility, class disparity.

The *ranking* is categorized into the high standard of living countries and low standard of living countries. People would like to choose the countries based on safety, resources availability, Individual facilities like travel services, employment, goods availability, communication facilities. People would like to live with affordable cost of pay for home, food, clothes, health issue treatments etc.

With the help of these three sections, the author of the thesis considering few main factors among standard of living and quality of life survey datasets. These datasets create awareness, confidence to the public who would like to find best places/countries for their own shelter, education, healthcare and business purposes in this world as per public choice. Each individual dataset is retrieved from the available world bank data source and required data information can be filtered or analysis. Datasets are collected in the form CSV or Excel format, that can be loaded into the chosen Tableau tool for further implementation and detailed visualization representation of data by applying diverse types of suitable pattern.

4.2 Workflow of the project

Fig -3 shows the workflow of the project. The author of the thesis created in the form of pictorial representation consists of survey datasets, task involved and working process of each task in the Tableau tool are involved in the project.

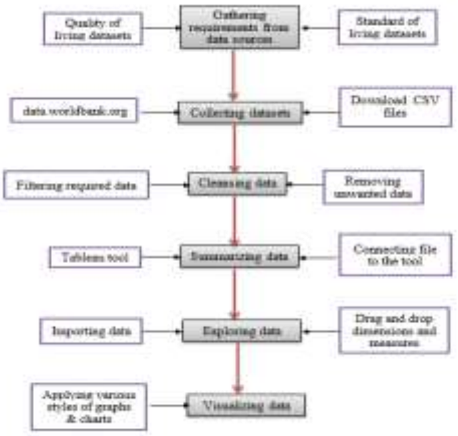

Quality of living datasets
Gathering requirements from data sources
Standard of living datasets
data.worldbank.org
Collecting datasets
Download CSV files
Filtering required data
Cleansing data
Removing unwanted data
Tableau tool
Summarizing data
Connecting file to the tool
Importing data
Exploring data
Drag and drop dimensions and measures
Applying various styles of graphs & charts
Visualizing data

Fig -3: Workflow of the project

5. DATA ANALYTICS PROCESS

For finding the best fittest places for the human survival, big data analytics techniques are involved to achieve final visualization report through the project. The process which is used in this project are:

1.Defining project specification

2.Defining requirements specification

3.Collection of datasets

4.Summarizing datasets

5.Exploring data In Tableau visualization tool

5.1. Visualization of quality of life

Table -2: Summary of comparison of quality of life results from the analyzed report

S.No	100% electrical accessibility countries	Best disaster reduction countries	Population above 10 lakh countries	Highly improved water source and sanitation facilities
1	Albania	Australia	Antigue and Barbuda	Andorra
2	Algeria	Bangladesh	Australia	Australia
3	Andorra	Barbados	Austria	Austria
4	Argentina	Brazil	Azerbaijan	Bahrain
5	Armenia	Canada	Bahrain	Belgium
6	Australia	Costa Rica	Bangladesh	Cyprus
7	Austria	Cuba	Belarus	Czech Republic
8	Azerbaijan	Ecuador	Belgium	Denmark
9	Bahrain	Germany	Benin	France
10	Barbados	Japan	Bosnia and Herzegovina	Germany

5.2. Visualization of standard of life

Table -3: Summary of comparison of standard of life results from the analyzed report

S.No	Literacy rates	Highly employment providing countries	Life expectancy
1	Andorra	Madagascar	Hong Kong and China
2	Azerbaijan	Qatar	Japan
3	Serbia	Uganda	Macao SAR, China
4	Spain	Rwanda	Italy
5	Venezuela,RB	Burundi	Spain

5.2. Representation of ranking countries

Table -4: Summary of high standard of living countries results from the analyzed report

S.No	1000 physicians available per 1000 people countries	Percentage of government health expenditure to the public	Percentage of individuals using the internet	Political stability and Non-violence countries
1	Austria	Andorra	Iceland	Greenland
2	Malta	Maldives	Liechtenstein	Singapore
3	Iceland	Nicaragua	Bermuda	New Zealand
4	Andorra	Marshall Islands	Bahrain	Liechtenstein
5	France	New Zealand	Andorra	Macao SAR, China

Table -4: Summary of low standard of living countries results from the analyzed report

S.No	Estimation of weakest governance performance	Mortality caused by road traffic injury(per 100,00 people)	Percentage of individuals using the internet
1	Syrian Arab Republic	Zimbabwe	South Sudan
2	Yeman, Rep	Venezuela,RB	Central African Republic
3	Afghanistan	Liberia	Sudan
4	Pakistan	Malawi	Chad
5	South Sudan	Congo, Dem Rep	Sao Tome and Principe

6. DISCUSSION

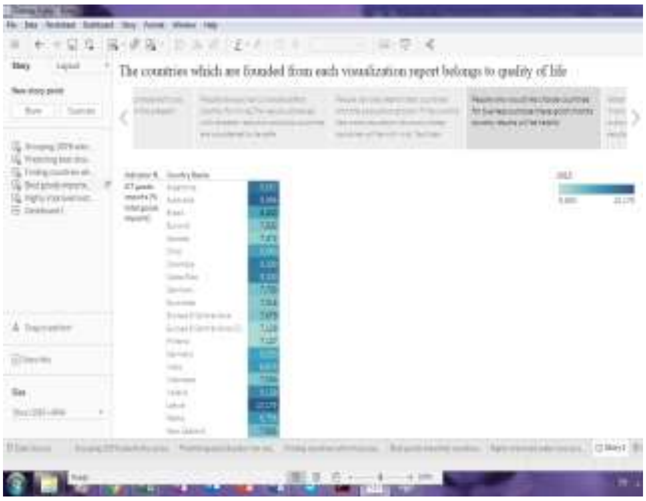
According to the goal of the thesis, the author of project comparers analysis results achieved from implementation of datasets into Tableau visualization tool based on the quality of life, standard of living to find out the best countries fit for public survival. In this section the author is discussing as follows:

a. How Tableau tool is responsive to the user and the great advantage of choosing this tool for the user.

b. How the public can decide the best countries with the help of the collection of analyzed report results using dashboard and story views.

Firstly, *section a*, description; Tableau tool is flexible for all users. There is no need of code or query implementation for achieving visualization results. Tableau tool has the built-in feature of executing queries automatically when datasets are connected to the tool. The author has chosen this tool because users without programming knowledge can work with this tool. For data analyst/data scientist can use this tool for achieving their business results.

Secondly, *section b,* description: The author of thesis says that all implemented datasets used for this project can be visualized at once in the Tableau tool. The story and dashboard command available in the Tableau tool helps user/data analyzer to compare all analyzed results together to make the better decision.



The countries which are founded from each visualization report belongs to quality of life

Fig -4: Story view

And also the user can bring out the solution based on the report result in a short period of time. Thus tableau tool reduces complexity in the visual representation. This tool allows the user to go forward and backward to view their results. The Fig 4.shows the visual representation of all the worksheet results together in the form of story creation.

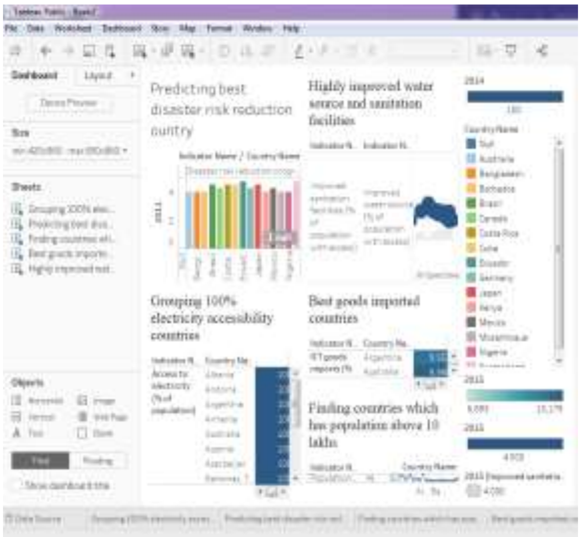
- The advantage of story narration helps the user to understand exactly what all are the data information available under each particular section.

- Story creation allows the users to add their own content about each visualization report for their future use.

The Fig 5 shows the visual representation of all the worksheet results together in the form of dashboard view:

- The advantage of dashboard view helps user to understand exactly what all the results achieved during analysis in the form of visualization report.

- Dashboard view allows the users to view the analyzed reports on their comfort device. This view is responsive to device types such as Phone, Desktop, and Tablet



Predicting best disaster risk reduction ountry

Highly improved water source and sanitation facilities

Grouping 100% electricity accessibility countries

Best goods imported countries

Finding countries which has population above 10 lakhs

Fig -5: Dashboard view

7. CONCLUSION

The author's concept of this project gives an idea to the people who are searching the best fittest country for the living, The author of thesis concludes the selection of best countries based on high standard countries and low standard countries. The public can also compare and analyze data information between high standard and low standard countries with poverty, literacy rates, healthcare facilities. In this project, analysis of datasets sre done on the basis of facilities which include infrastructure, environment, education, climate, political stability under two categories: quality of life and standard of living. In the summary of quality of life, standard of living and high standard ranking countries, the analyzed results obtained form each visualization report are compared with each other and matching countries are highlighted. These repetitive countries are considered as best countries and people can chose any of these countries for their survival. Thus the author concludes the project by saying that the achieved result of some countries which are found with the help of Tableau visualization representation by applying most important factors. Finally, the author compares the achieved results from the project with the Wikipedia results. Some of the countries Austria, Australia, Canada, Germany, Switzerland, New Zealand etc., found by analysis from the Tableau tool are matched with Global Livability Ranking, which contains list of most livable cities/countries are published every year by Economist Intelligence Unit(EUI).

REFERENCES

[1] Ali Syed Mohd., Gupta Noorpur, Nayak Gopal Krishan.,& Lenka Rakesh Kumar.," Big data visualization: Tools and challenges,"IEEE 2nd International Conference on Contemporary Computing and Information(IC3I),2016.

[2] Azevedo., Lourenco Ana Isabel Rojao., Santos., & Filipe Manuel, "KDD,SEMMA and CRISP-DM: a parallel overview" in ISCAP- Computers Communication in Scientific events,2008.

[3] Datasets used for the project from the web resource data world bank are available at https: // data. world bank .org/

[4] Downie Christopher Johnston.,& Peng Taoxin.,"Visualization of online datasets,", Software Engineering Research Management and Applications(SERA),IEEE 15th International Conference on,pp.,239-246,2017.

[5] Eri Thomas., Buhler Paul.,& Khattak Wajid.,"Big data adoption and planning considerations",2016,page 11.

[6] He Manchao., L Ribetro-e-Sousa.,& L Faramarzi.,"Rockburst process evaluation using experimental and Artificial Intelligence techniques,"1st Iranian Mining Technologies Conference, At Yazd,Iran,2012,Vol.24p

[7] M.D. Anto-Praveena & B.Bharathi., "A survey paper on big data analytics," in IEEE International Conference on Information Communication and Embedded Systems 2017,p9.

[8] M Vadovsky., P Michalik., I. Zolotova., & J Paralic., "Better IT services by means of data mining, "in IEEE 14th International Symposium on Applied Machine Intelligence and Informatics,2016,pp.187-192.

[9] Oussous Ahmed., Benjelloun Fatima-Zahra., Lahcen Ayoud Ait., & Belfkih Samir., "Big Data Technologies: A Survey, "Journal of King Saud University – Computer and Information Sciences,pp.,2017, ISSN 13191578.