Analysis of World Development Indicators

Kushal Dahal, Madan K C

1 Introduction

Many developing countries which were growing in the past decade are now developed. They are now economically stable. We have taken the consideration of economy, education, health and climate in this project. Conducting the exploratory analysis to understand how the countries are developing is our major plan. We want to classify the countries based on their economic development factors like GDP, GNI and Poverty. Also, visualizing how countries are raising their literacy rates is our second goal. To understand the health scenarios, we will work on people deaths due to communicable and non communicable diseases. We also want to explore the countries who are emitting CO2 the most. Other factor we have considered for development is how countries are growing their renewable sources as the non renewable sources are depleting each year.

For our motivation we want to analyze how the global development of the world in these two decades, we plan to work on development indicators that defines the core infrastructure development of a country. How countries are prioritizing economy, education, health, climate plays a significant role in the development of that particular country. It is interesting to understand how countries are dealing these aspects of the development. Country which are recently developed in this past decade can be an example for other developing countries to follow the path the developed one have gone through. On some countries, many people are dying because of communicable diseases due to their poor health situation. Some developed countries are growing their renewable resources but some developed ones are highly emitting Carbon Dioxide. We want to gain knowledge about these development aspects and visualize how countries are considering the global development.

2 Tasks

We will be working on five tasks that conduct the exploratory analysis on academic, health, economy and climatic development aspects of the world.

2.1 Literacy Rate and Government expenses on education overtime

We want to analyze the 'literacy rates' and 'Government expenses on education' of ten countries overtime from 2000 to 2020. We will consider five highly developed countries and five developing countries. How the literacy rates has changed in this past two decades? How the Government expenses have increased in this past two decades?

2.2 Countries with higher Carbon Dioxide Emission and air pollution

What are the countries that are responsible for emitting higher CO2? What are the countries with higher air pollution? Initially, we will sum the total CO2 emissions and air pollution of ten years (2010 to 2020) for each country. Further we will be identifying the countries with the higher values.

2.3 Proportion of total deaths due to communicable, non-communicable diseases and injuries

How people are dying due to following causes: communicable and non-communicable diseases and injuries? We want to analyze the proportion of total deaths due to those causes. Here, we will be considering the year 2019

2.4 Growth on production of electric energy from renewable sources

How the countries of Europe and North America have grown the production of renewable electric energy between two years (2015 and 2020)? We want to analyze the production growth and compare between these two continents.

2.5 Categorization of countries on the basis of economic factors

We want to categorize the countries on basis of economic factors like Poverty, GDP and GNI for year 2020. We will classifying them in three categories: developed, developing and least developed.

3 Literature Review

Berjawi [1] has worked on the hypothesis that men are more suicidal. He analysed the suicide mortality rate of males and females over the time from 1999 to 2019. From this, he found that the suicide mortality rate of males is almost 4 times greater than that of females and concluded his hypothesis as true. Later, the author analyzed the unemployment rate and found that the countries having higher unemployment rates have a higher rate of suicide mortality rate. Among them, the rate of suicide by males was higher than that of females. Moreover, he compared the suicide mortality rates with the countries with having high coverage of unemployment benefits and the countries having low coverage of unemployment benefits. There was a significant reduction in suicides as the coverage of unemployment benefits have increased. Finally, he concluded that suicides of males can be reduced by providing high coverage of unemployment benefits.

The journal by Qi Chen] [2] initially pictures the GDP of every country. He worked on the 'World Development Indicators' dataset published by the world bank. During the research, he compared two world maps of year 2000 and 2010 and how the GDP has changed for each country and also the change in government expenditure for healthcare has changed. Besides that, the author also demonstrated the relationship between life expectancy and GDP per capita for all countries. With the accomplishment of two tasks, the author has demonstrated that as the GDP increases, the expenses on the health sector of people also increases. Again, with the increase in GDP per capita, the life expectancy has also increased.

The article [6] by Jingwei Zhang, Yuming Wang, Lihong Feng, Changchun Hou, and Qing Gu had investigated the effects of air pollution and green spaces on impaired lung function in children. The authors tested the lung function of 2087 students of primary school in Tianjin aged between 9 and 11 years. They had evaluated the effects of indoor and outdoor environmental risk factors using the conditional logistic regression model. The risk of getting impaired lung function was increased by 53.4, 34.7, and 16.9 percent for every interquartile range increase in the mixture of six air pollutants at three lags: lag1, lag2, and lag3, and found that the protective effect of greenness at lag2 period was stronger. Though exposure to greenness had positive effects on lung health, most air pollutants were found to be hazardous to the lungs. As air pollution is rocketing day by day along with carbon dioxide emission, this article inspired us to visualize each country on this globe with the emission of air pollution and carbon dioxide emission.

• *E-mail: dahalku@mail.gvsu.edu, kcma@mail.gvsu.edu*

The article [5] by Wilson X.B. Lia and Tina T. Hea clarified that external assistance should be provided to nonviable residents to solve the poverty issue. The authors proposed a model to exhibit the involvement of government in eradicating poverty. With the concept of individual and social resources, they defined a mathematical function of fighting poverty satisfying Cobb Douglas function and explained how public market mechanisms mobilizing societal resources are necessary along with strong state capacity, competent leadership, and high social trust. Furthermore, the authors applied the model and concept of viability to compare the war on poverty in US and poverty alleviation in China. They also conducted a cross-country investigation to obtain empirical evidence which supported this model and concept. Overall, this article inspired us to categorize countries into developing, developed, and least developed on the basis of poverty, GDP, and GNI of each country.
Another useful article [4] by Jyothi Kumbar, Neelaganga Goudar, Sahana Bhavi, Shruti Walikar, and Basanagouda F. Ronad which concluded that the renewable energy sources can be effectively employed for energizing the hospital with grid-connected mode. The authors presented the HOMER simulation model to optimize renewable energy system components which were conducted for 100 beds and 50 beds COVID-19 hospital located in Bagalkot city, Karnataka, India. The solar energy potential solar PV panels in the range of 10-100 kW for 100-bed hospitals and 10-50 kW for 50-bed hospitals were employed. They also collected wind speed data from the weather stations to use as input resources for wind turbines. They developed two simulation models for each hospital and supervised the optimization of energy system components based on the cost of energy and net present cost. All possible configurations were simulated in HOMER and ten feasible options were listed which showed that the grid-connected system was found to be the most economical and reliable option for the electric energy loads. Hereby, this article gave us the idea to analyze the production of renewable energy and comparison of the production between two continents: North America and Europe.

4 Overview of Data

We got the 'World Development Indicators' dataset [3] of size 57Mb containing details of all countries with 1574 types of development indicators presenting the core infrastructures scenarios of each year from 1960-2020 (Health, Education, Employment, Economy, Transportation, Climate, Water and Electricity, Pollution, Population, Violence). There are 383k rows and 65 columns. The countries are also classified into continents. World Development indicators are the combination of all aspects of development compiled by the world bank. Among the 1574 development indicators, we will be working on 12 significant indicators which can visualize the economy, academic, health, climate aspects of the global development. We will be considering the data from 2000 to 2020.

- Government expenditure on education, total (Percentage of government expenditure)
- Literacy rate, adult total (Percentage of people ages 15 and above)
- CO2 emissions (Kiloton)
- PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)
- Cause of death, by communicable diseases (Percentage of total)
- Cause of death, by injury (Percentage of total)
- Cause of death, by non-communicable diseases (Percentage of total)
- Poverty headcount ratio at national poverty lines (Percentage of population)
- GNI (current US Dollar)
- GDP (current US Dollar)
- Electricity production from renewable sources, excluding hydroelectric (Percentage of total)
- Electricity production from hydroelectric sources (Percentage of total)

5 Project Design

5.1 Literacy Rate and Government expenses on education overtime

In this task, we want to visualize how the literacy rate and government expenses have changed over these past two decades over time. We would like to compare these two changes between developing and developed countries. We have considered five developing and five developed countries for this task. To conduct the task, we will be building two line graphs each one for Literacy Rate and Government expenses. Line graph can be beneficial and visualizes the overtime of data in a simple manner that will be easier to understand. The line graph is very simple and most interpreters are used to seeing and understanding this graph which fulfills the law of isomorphic correspondence.
For this task, a scatterplot and bar chart can also be used. If we had considered scatterplot, there would be dots everywhere. The difficulty would be to show the overtime of years simply as compared to the line chart. For the bar chart, there would be the bars of 20 years each year having five developed and five developing. This would make the chart very complex to analyze and the problem of scalability affects.
For the line chart, position and color are the visual encodings. For the position visual encoding, the x-axis of the line chart will represent the year and the y-axis represents the government expenses and Literacy rate. All the ten lines of developed and developing countries will show the positions on each instant based on the x and y-axis values. Cleveland's rules also list the position as the most accurate visual encoding as it accurately shows the positions at each instant of the line chart. This will make the interpreter accurately understand the graph. Also, countries with similar positions have similar kinds of characteristics fulfilled by the Law of proximity.

In the line chart, color is another visual encoding that is implemented. Color is used to separate the developing and developed countries. Here we have used a green color line to show the five lines of developed countries. The green color is chosen as it shows the positive vibes for the interpreter and developed countries have higher literacy and government expenses rate. Also, developing countries are slowly increasing their literacy rates and government expenses which can be interpreted easily by the blue lines as blue color creates calm, slow, and aloof vibes for the interpreter. Cleveland's rules also support the identification of lines in developed and developing countries using color encoding. The Gestalt Law of similarity shows the countries with a similar color. Gestalt Law of focal point explores the countries which are green showing positive growth.

5.2 Sum the CO2 emissions and air pollution for ten years (2010 to 2020) for each country and identify the countries with the higher values

Intending to visualize the total CO2 emissions and total air pollution by summing the data for ten years from 2010 to 2020, we want to explore the countries emitting the most CO2 and high air pollution around the world. We will be building a geospatial world map with countries divided by the border. As we have been looking at the world map since childhood, this map is familiar and easier to interpret for all people. The world map will show the position and area of that particular country in the world. Gestalt's Law of isomorphic correspondence also shows that all countries are in the place they used to be on a standard map and most interpreters are used to seeing it. Again the viewers can easily separate the countries from one another with the help of borders fulfilling the Law of the enclosure.

Graphs like bar charts and treemaps can be used for this task but we did not implement them here. While we can sort the bars based on air pollution, it will be difficult to find specific countries or to find

the countries with the highest CO2 emission. Our task seeks to allow users to identify specific countries that have the highest air pollution or the countries that have the higher CO2 emission with ease. With a bar chart, it also will be difficult to consider all 200+ countries in one graph. The bar chart is not very scalable. Again for a treemap, implementing 200+ countries in one graph is very difficult to interpret. The area of the treemap and the color density of each instant can show the CO2 emissions and air pollution. Though we can find the countries with high air pollution and CO2, it will be difficult to find specific countries and know the details. When we compare these two graphs with our world map graph, the world map graph is easier to interpret and is familiar and also the positions of the countries are well-displayed.
For this graph, visual encodings such as color, area, and position will be implemented. We have implemented position encoding to show the place each country inherits in the world. As Cleveland's rules have also supported that position has more accuracy to visualize the country's location where they are placed in the geographical region, it will be easier for the interpreter to understand the graph. While looking at the Law of proximity, the countries near each other have the same kind of characteristics.
We will use area visual encoding for two purposes. One is to show how much of the world each country occupies. Another purpose is to explore the CO2 emission by implementing the area of a dot with dark blue color. When the countries have higher CO2 emissions, the area of the dot will increase. Cleveland supports the area which is used to classify different countries and different CO2 emissions. Here Gestalt Law of similarity is also applicable as countries can have similar areas of the dot indicating similar CO2 emissions.
We will use color encoding implementing sequential color scales to visualize the total air pollution for each country. For the sequential color scale, we will be choosing the red spectrum. As the air pollution increases, the darker the country's color becomes. We have chosen red because as the countries become darker red, the air pollution increases which is a negative factor and hazardous for health. Gestalt Law of focal point explores the countries which are more dark red and have higher attention. This will help the interpreter easily feel like the country with darker red color has more air pollution which can affect people's health and also the dots with large areas have higher attention. Again Gestalt Law of similarity shows the countries with a similar color. Cleveland's Rule also supports the identification of countries with higher air pollution using color encoding. We have chosen dark blue because this color is easily recognizable having a red background so the Gestalt Law of focal point explores the countries which have a higher area of dots and have higher attention.

5.3 The proportion of total deaths due to communicable, non-communicable diseases and injuries

We want to visualize the proportion of total deaths due to communicable, non-communicable diseases and injuries and identify the case with the highest proportion using a pie chart. We chose the pie chart as we know most of the interpreters are very familiar with the pie chart since school and this chart can easily be analyzed by the nontechnical audience too. Gestalt's Law of isomorphic correspondence also shows interpreters are used to seeing the standard pie chart. Again for the viewers, the proportions of each case can be easily separated from one another with the help of borders in the pie chart fulfilling the Law of the enclosure.
Graphs like treemap and doughnut charts can be also used as alternatives. But the treemap can be favorably used for more than three cases as pie charts might work best for three or few segments and interpreters are more used to seeing the pie chart rather than the heat map. Again when I compared doughnuts with the pie chart, the pie chart easily shows the immediate understanding of proportions with higher visual efficiency. Also, interpreters are more used to seeing the pie chart rather than the doughnut. So I chose a pie chart among these graphs.
Color and area are the visual encodings I have implemented in this task. The area shows the proportion of total deaths caused by each case like communicable diseases, non-communicable diseases, and injuries. Cleveland supports the area which is used to visualize the proportions for the interpreters effectively. Here Gestalt Law of similarity is also applicable as two cases may have similar areas representing two cases with almost equal proportions.
We will use color encoding implementing categorical color scales to visualize the proportion of deaths due to different causes. For the categorical color scale, we will be choosing red, yellow, and orange. We will choose the red color for the cause of deaths with a higher proportion because the red color seeks higher attention. Also, we have chosen yellow and orange colors because these colors show the caution and unsafe feelings for the interpreters as the whole chart represents the proportion of deaths. Cleveland's Rule also supports the identification of causes of death with a categorical color scale using color encoding. Gestalt Law of focal point explores the causes of deaths with a higher proportion using red color and have higher attention.

5.4 Growth in the production of electric energy from renewable sources

In this task, we are comparing the production of renewable electric energy across North America and Europe and comparing between two years 2015 and 2020. We are using a bar chart to perform this task. It will be easier for us to compare the length of bars across these two scenarios. The bar chart is the most simple graph creating easier understanding for the interpreters to understand, fulfilling Gestalt's Law of isomorphic correspondence as they are used to seeing standard bar charts. Again the graph is very simple with uncomplicated visualization so Gestalt's Law of Pragnanz is also fulfilled.
We could have also used a Scatter plot and Line chart for this task. Though a Scatter plot can visualize each data point with the dots, it will be difficult for the interpreter to compare the electric energy between each year and continent. The accuracy of the efficiency to study the graph will certainly decrease for the scatterplot. For the line chart, though the chart will show the growth of electric energy between two years, it will be difficult to identify and compare the accurate differences in electric energy production. While comparing the Bar chart with these two alternative graphs, a Bar chart will help interpreters accurately understand the difference between those two years and between those two continents.
Position, Length, and color are the visual encodings implemented in this task. For the position encoding, the x-axis of the bar chart shows the year, and the y-axis represents the electric energy production. Cleveland also supports that interpreters more accurately identify the position encoding. This bar chart also has Length encoding. The length of the bars represents the renewable energy production values. We can easily see the trends of bars because of the length of the bars. Again, the Clevelands rule supports that length can maximize the accuracy of recognition of the graph.
We will use color encoding to separate the continents. Each color represents one particular continent. Above we have implemented a categorical color scale in the graph to separate the two continents. We have chosen green and blue colors for two continents as these two colors give a positive feeling to the interpreters. The growth of production energy is a positive thing. So these two colors can show the positive aspect. The implementation of a categorical color scale to separate two continents is also supported by the Clevelands rule. The Gestalt Law of similarity also applies as two bars can have the same colors.

5.5 Categorization of economy of countries based on GDP and Poverty

We are building a scatterplot to categorize all countries of this world based on their economic factors like GDP and poverty and visualize the world economic situation with a graph. We know that countries with higher GDP and lower poverty are more developed than countries with lower GDP and higher poverty. In this task, we will be building three clusters, each cluster representing developed, developing, and least developed countries based on their GDP and poverty. To build these three clusters, we will be implementing a k-means clustering algorithm. We will be building three groups using this unsupervised clustering process. The size of the clusters will be easier for the interpreters to see how many countries of this world are more developed and compare

with the developing and least developed countries.
We could have used a bar chart and bubble chart in this task. The length of the bars of a bar chart could have shown the numeric attributes, but as we have considered 206 countries, there will be 206 bars. So it will be difficult for the interpreters to categorize which countries are developed, developing, and underdeveloped countries. Again it will be difficult to see how many countries are more developed, least developed among 206 countries. For the bubble chart, there will be a lot of overlap between the data points and it will be difficult to accurately understand the clusters with more data points.
In this task, visual encodings position and color are used to build the graph. Here, the x-axis represents the GDP of a country and the y-axis represents the poverty of a country. Cleveland also supports that with the help of the position of each data point, it will be easier for the interpreters to accurately identify clusters at which positions are more developed and which ones are least. Again two countries can have similar positions which fulfill Gestalt's law of similarity. For Gestalt's law of proximity, we can see that the points in the scatter plot which are near to each other can be grouped. For the law of closure, while looking at the scatter plots, my mind starts to set the clusters for the points having similar kinds of GDP and poverty.
We have used color to categorize all the countries of the world in three clusters. Each cluster has a particular color. A categorical color scale is implemented in this task. Cleveland's Rule also supports the identification of countries grouped in one cluster using color encoding. The cluster which is highly developed is represented by green color because it shows the positive approach for the viewers. The cluster representing developing countries is represented by the yellow color as it shows calm and neutral perception. Again the red color is used to show clusters of least developed countries as the red color provides negative feelings to the interpreters. Gestalt's law of focal points also fulfills this task as the red color is implemented for the least developed countries cluster. Gestalt law of similarity fulfills that two countries in a cluster can have the same color.

6 Visualization and Analysis

6.1 Literacy Rate and Government expenses on education overtime

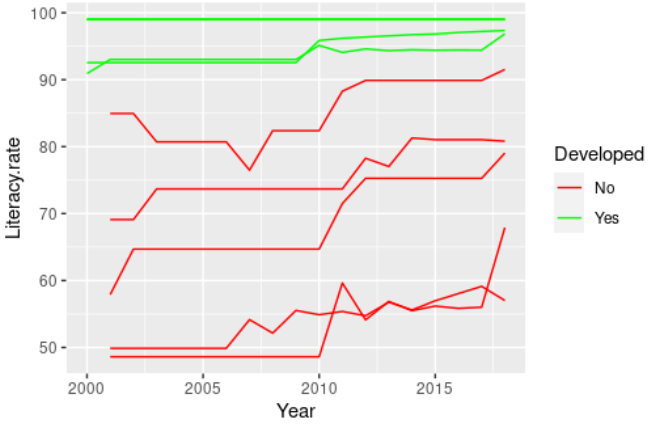

Fig. 1. Line Chart of Literacy Rate

Our first task was to visualize the trend of literacy rate and the government's expenses on education over the time from 2000 to 2020. As per the design phase, we have used two line charts to find the trend of literacy rate and government expenses for the education of five developed countries and five developing countries.
We have considered the United Kingdom, Switzerland, China, Australia, and Singapore as developed countries whereas Nepal, Ghana, Guatemala, Namibia, and Pakistan as developing countries. The government expenditure was taken in the percentage of total expenditure by each country and the literacy rate was also measured in the percentage

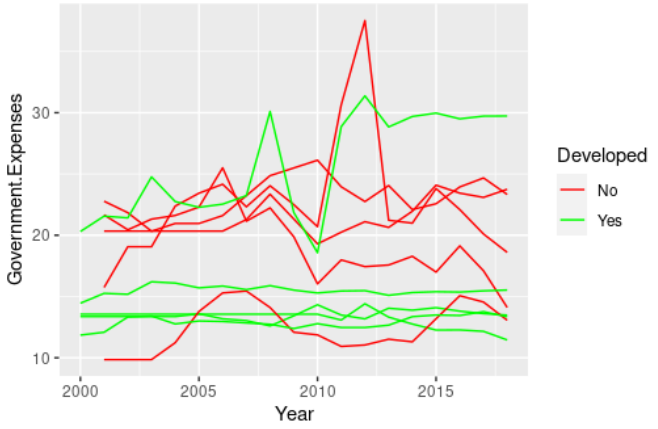

Fig. 2. Line Chart of Government Expenses

of the people above 15 years. We have used color encoding here by drawing lines of developed countries in green color and the developing countries in red color supposing that the developed countries have higher literacy rates and expend more on education.
We visualized line charts in R where we found that all the developed countries have higher literacy rates than the developing countries exceeding 90 percent in each country. Similarly, the trend shows that the literacy rate is inclining upward each year. In the case of government expenditure, we visualized that majority of the developed countries have higher expenditure on education but there is no large difference in the expenses between developed and developing countries in literacy rate. We can also visualize that the trend of government spending on education is fluctuating in most countries.
We had chosen the line charts instead of scatterplots and bar charts due to the reason that line charts can visualize the overtime data in a simple and easily understandable manner which is justified by this graph as well. Furthermore, the position of each line representing countries has increased the accuracy of the visualization as per Cleveland's Rule. We visualize that the countries having positions near to each other have similar literacy rates and government expenses in both charts also support the Gestalt's law of proximity. Another visual encoding called the slope of the lines has also helped to visualize accurately. Although the color encoding is not considered as accurate as position and slope, it definitely adds some accuracy to charts and we can visualize that lines having the same color are similar to each other which supports the law of similarity. Furthermore, we can see Gestalt's Law of focal point being applied where the red color is visualized at the first sight.

6.2 Identify countries with the higher values of CO2 emission and air pollution in world

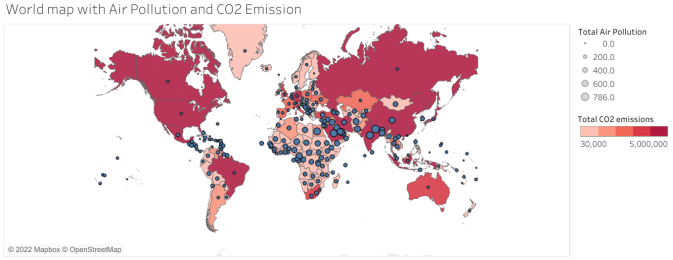

Fig. 3. World map with higher CO2 Emission and Air Pollution

The world map above is the visualization of our task 1 i.e. identifying the countries with the higher value of Carbon dioxide emissions and air pollution for ten years from 2010 to 2020. The map effectively visualized both carbon dioxide emission and air pollution in each coun-

try separately.
The carbon dioxide emission was in the unit 'Kiloton' and the air pollution in 'micrograms per cubic meter'. We summed up the carbon dioxide emission in all the countries across the globe for the 10 years period and also summed the air pollution in that period of time. We used the sequential color scale of the red spectrum to visualize the total carbon dioxide emission. The darker red color of a country shows that the country has higher emissions of carbon dioxide and the light red color shows lower emissions. We visualized the air pollution by the blue circle in each country where the size of the circle in each country increases when the air pollution of that country is higher.
Using Tableau for the visualization, we identified the countries with carbon dioxide emissions in red color scale and air pollution by the circle. The world map shows that the North American countries, most of the Asian countries, and other countries such as Brazil have higher Carbon dioxide emissions. And most of the African countries and a few other countries like Greenland have fewer emissions. While observing the air pollution level, we can visualize that most Asian and African countries have more air pollution whereas American and European countries have less air pollution. One interesting finding we found in this visualization is that initially, we thought the country having higher carbon dioxide emissions would also have a higher level of air pollution. However, the two terms are not higher together in any country except the Arabian and South Asian countries.
We had chosen the world map instead of the treemap and bar graph in the design phase thinking that the world map is familiar to people and easier to understand. The decision is found to be right, and the position of each country on the world map, the area occupied by countries, the area of the circle to show the air pollution, and the red color to show carbon dioxide emission have also supported Cleveland's Rule to increase the accuracy of the visualization. And the position of countries similar to the standard map has also been supported by Gestalt's Law of Isomorphic correspondence. Moreover, we had chosen red color, which shows the negative emotion, to visualize the carbon dioxide emission as it adversely affects the globe. And when we visualize the map, we can see the countries with red color in the first sight which is also supported by Gestalt's Law of the focal point. The way we used blue color for the circle in each country having the red color spectrum in the background is also easier to recognize. Besides these all, we can visualize that countries having the same darkness of color have similar values of carbon dioxide emission which also supports the Gestalt principle of similarity.

6.3 The proportion of total deaths due to communicable, non-communicable diseases and injuries

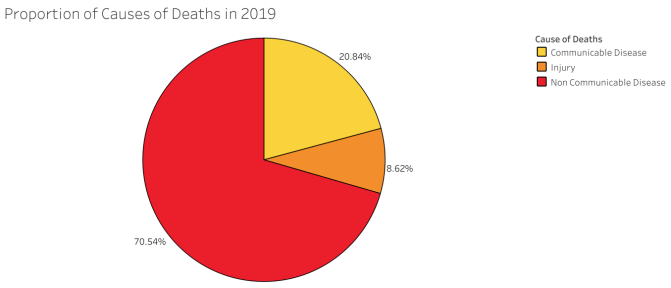

Fig. 4. Pie chart with proportion of cause of Deaths

Our third task is to analyze the proportion of total deaths due to three causes: communicable diseases, non-communicable diseases, and injuries. As per the design, we effectively visualized the proportion of these causes in all countries for the year 2019 in a pie chart. To visualize the proportion of the total deaths, we summed each value of the causes of each country and then visualized the pie chart in Tableau. After the completion of the pie chart, we found Non Communicable diseases to be more responsible for death with 70.54 percent of total deaths. Again 20.84 percent of total deaths in this world is caused by communicable disease and only 8.52 percent of total deaths are caused by injuries. Hence, after analyzing the pie chart, we came to the conclusion that non-communicable disease indicated by the red proportion shows it is highly responsible (70.54percent of total deaths) for the deaths of the people in this world.
While working on the design phase, we decided to build the pie chart after comparing it with the other few alternatives chart. As Gestalt's principle Law of isomorphic has shown that the interpreter is used to seeing the standard pie chart. After the final pie chart, we felt that it is true and anyone can easily understand the chart. Similarly, the Gestalt principle of Law of the enclosure has also supported the separation of proportions of each cause inside the pie chart with the help of borders. During the design phase, we had chosen the red, yellow, and orange colors thinking that the red color will play a significant role to make the interpreter have negative emotions while looking at the chart. After the completion of the pie chart, the red proportion indicating non-communicable diseases causing the highest amount of deaths can be easily visualized at the first sight. So the Gestalt law of focal points supports the pie chart. Again choosing the categorical color scale for this graph has helped to distinguish the proportions that also support Cleveland's Rule as the accuracy in the visualization increased in the chart due to color encoding. Moreover, all three colors show negative emotions which accurately visualize the death of people.
Similarly, each cause has covered a certain area and angle of the pie chart. Cleveland's rule also supports area and angle for the accuracy of the pie chart. We also felt that we did the right decision to choose the pie chart to find the proportion of three causes after comparing it with other charts because this pie chart has shown the proportions in a simple, understandable, and accurate manner which has made the interpreter easier to understand. So the decisions like choosing the red color, choosing a pie chart among other charts, implementing Gestalt laws and Clevelands rules, and choosing a categorical color scale have played a significant role to accomplish this task by building an effective pie chart visualizing all the requirements of the task in a simple and most understandable manner.

6.4 Growth in the production of electric energy from renewable sources

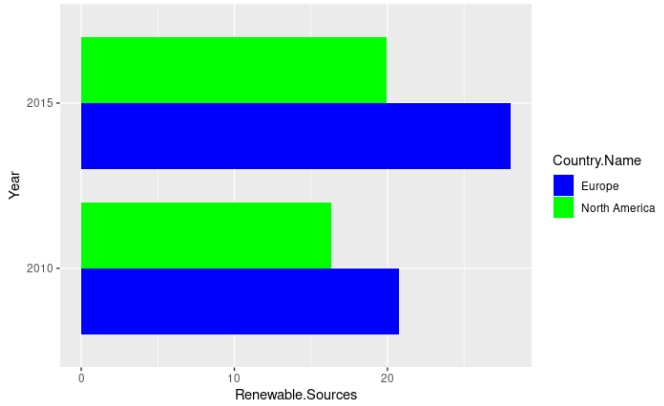

Fig. 5. Bar Chart of electric energy from renewable sources

Our fourth task was to compare the production of renewable electric energy in North America and Europe in the two years 2015 and 2020. The horizontal bar chart successfully visualized the comparison of the two continents in the two years which is shown in the screenshot above. While doing this task, we took two attributes: 'Electricity production from renewable sources, excluding hydro-electric' and 'Electricity production from hydro-electric sources', both in the percentage of the total. Then, we added these two percentages per year and then visualized them in bar charts using R where percentages of energy production in the continents were kept on the y-axis and years(2015 and 2020) on the x-axis respectively.

We had planned to use a bar chart instead of a scatterplot and a line chart due to reason that the comparison of only two continents in a year is easier to visualize and compare accurately than any other graphs. We can see that the position of the bars which is also considered the most accurate visual encoding has helped this bar chart to visualize more accurately. The length encoding has also increased the accuracy and we can compare the difference in the percentage of energy production between these continents easily. The use of two categorical color scales to separate the two continents, green for North America and Blue for Europe has also increased the accuracy as we can easily compare between two continents. The green and blue colors are positive ones and the energy production is also positive and we can visualize the positivity in the production. We visualized that the bars having similar colors are similar to each other and in this graph, we separated continents using the color scale which is also supported by Gestalt's principle of similarity. Besides that, we also visualized the horizontal graphs stacked together to have a common value i.e. year on the x-axis which also justified Gestalt's law of proximity.
From this graph, we can easily visualize that the production of energy is higher in Europe in both the years. Moreover, the percentage of the production in North America in the year 2015 is near to that in Europe in the year 2020.

6.5 Categorization of the economy of countries based on GDP and Poverty

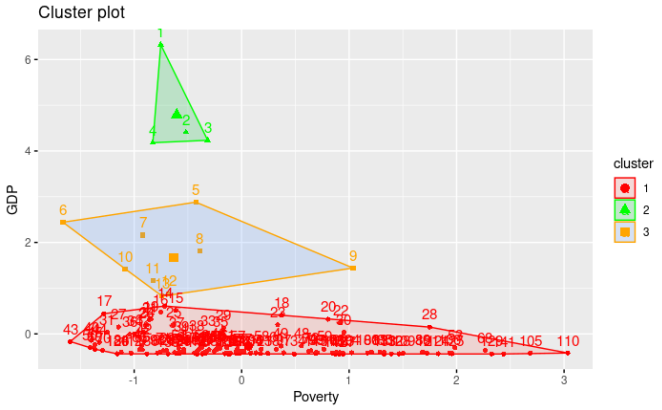

Fig. 6. Categorization of Economy of Developed and Developing Countries

Our aim in the fifth task was to categorize the countries based on GDP and poverty. As per the design phase, we used a scatterplot to categorize all countries in the world into three clusters(developed, developing, and underdeveloped) implementing the k-means algorithm on the basis of GDP and poverty rate in the year 2020. We have taken the attribute 'Poverty headcount ratio at national poverty lines' and GDP from the data source. Poverty was in the percentage of the total population of each country and the GDP was in US Dollars. Using R, we visualized the scatterplot with three clusters by adding the poverty on the x-axis and GDP on the y-axis.
As per the design, we chose scatterplot over bar chart and bubble chart thinking that more than 200 countries and their clustering into three categories can be visualized accurately and easily only in the scatterplot. From the scatterplot using library ggplot2 in R, we categorized the countries as developed, developing, and underdeveloped and found this to be easily understandable and accurate. The position of each country in the scatterplot has supported Cleveland's Rule. With the help of the position of the countries in the plot, clusters were easily created. The countries near to each other have been grouped into a cluster i.e. developed, developing, or underdeveloped which also supports Gestalt's law of proximity. Similarly, we can see that the shape of each country in each cluster is common which increased the accurate categorization and this also supports the Gestalt's Law of similarity. Furthermore, the categorical color scale has also increased the accuracy of the scatterplot here. When we visualize the plot, we visualize the cluster with the red dots at first which also supports Gestalt's law of focal point. As the red color has negative emotion, we easily understood from the scatterplot that the cluster having red color are underdeveloped countries having lower GDP and higher poverty.
From the scatterplot, we found that most of the countries are in the red cluster which means that these countries are underdeveloped and have low GDP and high poverty. Few countries are in the orange cluster of developing countries and very few countries are in green cluster of developed countries.

References

[1] Z. Berjawi. Men are more suicidal. American University of Beirut, November 2021.

[2] Q. Chen. Data visualization and analysis, part 1/3 – world bank indicator, January 2016.

[3] W. B. Group. World development indicators, 2022. https://datacatalog.worldbank.org/search/dataset/0037712.

[4] J. Kumbar, N. Goudar, S. Bhavi, S. Walikar, and B. F. Ronad. Optimization of hybrid renewable electric energy system components for covid hospitals. In *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pp. 799–804. IEEE, 2021.

[5] W. X. Li and T. T. He. Political order and poverty eradication. *Frontiers of Economics in China*, 16(3):470–494, 2021.

[6] J. Zhang, Y. Wang, L. Feng, C. Hou, and Q. Gu. Effects of air pollution and green spaces on impaired lung function in children: a case-control study. *Environmental Science and Pollution Research*, 29(8):11907–11919, 2022.