Interactive Visualization of World Bank – Global Financial Development Dataset

Abstract

This project is about presenting a humungous dataset in the form of interactive data visualization. Data visualization is the representation of data or information in a graph, chart, or another visual format. It communicates the relationships of the data with images. Interactive Data Visualization is important because it allows trends and patterns to be more easily seen. We aim to assimilate appropriate and usable interactive data visualization techniques. In this project, we will take advantage of available data by putting it into interactive data visualizations, which can reap various benefits for the user while manipulating the data to find out specific things that they need to know.

Introduction

Data visualization has changed our society considerably. From a most simple projected line across a football field through to complex graphs outlining market fluctuations, they are changing the way that our society is approaching and understanding data. However, despite the huge impact visualizations have had, they still face considerable challenges in the future. Augmented reality may well be the single biggest change that we are going to see regarding the use of data visualizations. Virtual reality is going to have a huge impact on the potential for data visualizations, allowing people to interact with data in the third dimension for the first time. As we move toward more interactive and complex trends for data visualizations, we are going to be seeing an increased need for technical skills to first understand and translate the data then create visualizations around the results. Convenient and effective visualization hence becomes a necessity to harness the true power of data analysis. The major area we hope to cover here is the hierarchical form of data since there have been few techniques developed for the visualization of hierarchical form of data but all of them have had certain shortcomings, which we hope to overcome.

Interactive Visualization of World Bank – Global Financial Development Dataset

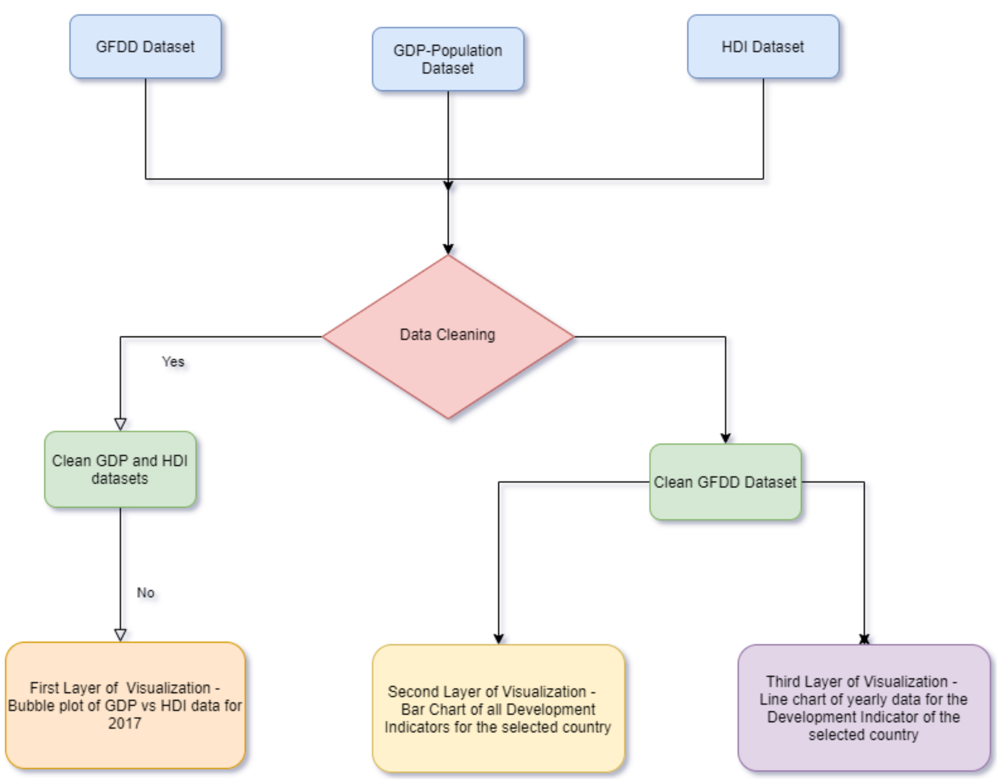
Objectives of the study

The overall objective of the present study is to create various levels of visualizations and limit the visualization to one level at a time. The specific objectives are:

- The 1st level of hierarchy is a bubble plot visualized from GDP and HDI datasets.
- The 2nd level of visualization is from the select country to all its corresponding factors.
- The 3rd level of visualization is from the development indicator to all the data available of that particular indicator.

Methodology

Flow Chart:



Interactive Visualization of World Bank – Global Financial Development Dataset

M.1 Data collection: The main data has been used from the World Bank website. This data is about the Global Financial Development (GFDD) over the last 15 years. The dataset includes the primary dataset used for visualization, and the other two datasets shown below are taken from other websites (links given in references) :

	A	B	C	D	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI
1	Country Name	Country Code	Indicator Name	Indicator Code	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2	Afghanistan	AFG	5-bank asset conce	GFDD.OI.06			100	100	100	89.3635	83.133	86.6647	84.8274	79.6688	86.6035	72.1549	71.9406
3	Afghanistan	AFG	Account at a forma	GFDD.AI.05								9.00501			9.961		
4	Afghanistan	AFG	Account used for b	GFDD.AI.08								2.7184					
5	Afghanistan	AFG	Account used to re	GFDD.AI.09													
6	Afghanistan	AFG	Account used to re	GFDD.AI.10											2.25439		
7	Afghanistan	AFG	Account used to re	GFDD.AI.11											2.54747		
8	Afghanistan	AFG	ATMs per 100,000 a	GFDD.AI.25	0.015509	0.059468	0.115723	0.205091	0.297777	0.446217	0.525093	0.611646	0.632651	0.700559	0.744878	0.911542	1.05214
9	Afghanistan	AFG	Bank accounts per	GFDD.AI.01					36.1814	86.9145	105.115	138.489	165.019	154.763	171.692	179.961	180.806
10	Afghanistan	AFG	Bank branches per	GFDD.AI.02	0.364464	0.579815	0.918548	1.23054	1.46811	2.2108	2.36948	2.20066	2.12089	2.23949	2.2894	2.14789	2.09421
11	Afghanistan	AFG	Bank capital to tota	GFDD.SI.03													
12	Afghanistan	AFG	Bank concentratioi	GFDD.OI.01		100	97.3211	89.1844	92.7161	73.0637	59.7381	66.5181	63.8239	57.8009	63.7336	53.3791	51.8292
13	Afghanistan	AFG	Bank cost to incom	GFDD.EI.07	95.9831	62.0513	61.6096	94.5866	58.7023	59.0729	64.3447	80.1512	75.9846	75.4972	75.862	72.21	56.3961
14	Afghanistan	AFG	Bank credit to banl	GFDD.SI.04			52.9056	56.8219	59.8813	58.2675	59.4385	43.1852	24.4893	23.3185	22.6016	21.8267	20.9612
15	Afghanistan	AFG	Bank deposits to G	GFDD.OI.02			8.99636	9.92974	13.7797	15.3914	17.1166	17.6887	16.5873	17.1107	16.911	16.4736	16.6096
16	Afghanistan	AFG	Bank lending-depc	GFDD.EI.02													
17	Afghanistan	AFG	Bank net interest r	GFDD.EI.01		9.79808	9.86467	8.21465	10.3925	11.0339	5.59061	6.44951	5.89362	6.19178	5.26145	5.00389	4.94282
18	Afghanistan	AFG	Bank noninterest i	GFDD.EI.03	43.8819	47.1795	39.0535	20.9122	15.7293	21.3292	42.8146	29.5158	48.6004	40.8048	42.0813	46.5658	61.8637
19	Afghanistan	AFG	Bank non-perform	GFDD.SI.02													
20	Afghanistan	AFG	Bank overhead cos	GFDD.EI.04		7.6925	7.55613	7.62319	5.49604	6.03932	3.479	4.52654	4.20954	3.8448	3.62373	3.36178	3.15006
21	Afghanistan	AFG	Bank regulatory ca	GFDD.SI.05													
22	Afghanistan	AFG	Bank return on ass	GFDD.EI.05		2.01434	1.64735	-0.46867	2.14344	1.23836	0.766775	-0.61966	0.284938	0.56277	0.466695	-0.01057	1.56259
23	Afghanistan	AFG	Bank return on ass	GFDD.EI.09		2.33736	2.47054	-0.54662	2.75981	1.6049	0.798694	-0.61505	0.584539	0.810591	0.605692	-0.02195	1.93454
24	Afghanistan	AFG	Bank return on equ	GFDD.EI.06		7.09843	5.61671	-1.94585	15.0635	8.766	4.89186	1.55755	11.3692	8.26648	6.64126	0.068273	14.8124
25	Afghanistan	AFG	Bank return on equ	GFDD.EI.10		8.23672	8.42339	-2.26949	19.3952	11.3606	5.09549	1.59597	14.6359	10.8612	8.11002	-0.05188	18.3383

Figure 1.1 Raw dataset from the World Bank

	A	B	C	D	E	F
1	Entity	Code	Year	Population density (people per kmÂ² of land area)	GDP per capita (constant 2011 international $)	Total population (Gapminder)
2	Afghanistan	AFG	1800			3280000
3	Afghanistan	AFG	1801			3280000
4	Afghanistan	AFG	1802			3280000
5	Afghanistan	AFG	1803			3280000
6	Afghanistan	AFG	1804			3280000
7	Afghanistan	AFG	1805			3280000
8	Afghanistan	AFG	1806			3280000
9	Afghanistan	AFG	1807			3280000
10	Afghanistan	AFG	1808			3280000
11	Afghanistan	AFG	1809			3280000
12	Afghanistan	AFG	1810			3280000
13	Afghanistan	AFG	1811			3280779
14	Afghanistan	AFG	1812			3282342
15	Afghanistan	AFG	1813			3284692
16	Afghanistan	AFG	1814			3287834
17	Afghanistan	AFG	1815			3291770
18	Afghanistan	AFG	1816			3296506
19	Afghanistan	AFG	1817			3302044
20	Afghanistan	AFG	1818			3308390
21	Afghanistan	AFG	1819			3315547
22	Afghanistan	AFG	1820			3323519
23	Afghanistan	AFG	1821			3332311
24	Afghanistan	AFG	1822			3341926
25	Afghanistan	AFG	1823			3352368

Figure 1.2 Raw GDP dataset

Interactive Visualization of World Bank – Global Financial Development Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	HDI Rank (2018)	Country	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	200
2	170	Afghanistan	0.298	0.304	0.312	0.308	0.303	0.327	0.331	0.335	0.339	0.343	0.345	0.347	0.378	0.387	0.4	0.41	0.41
3	69	Albania	0.644	0.625	0.608	0.611	0.617	0.629	0.639	0.639	0.649	0.66	0.667	0.673	0.68	0.687	0.692	0.702	0.70
4	82	Algeria	0.578	0.582	0.589	0.593	0.597	0.602	0.61	0.619	0.629	0.638	0.646	0.655	0.666	0.676	0.685	0.694	0.69
5	36	Andorra	..	..	..	..	..	..	..	..	..	..	0.759	0.767	0.78	0.82	0.826	0.819	0.82
6	149	Angola	..	..	..	..	..	..	..	..	..	0.384	0.394	0.404	0.419	0.428	0.44	0.453	0.46
7	74	Antigua and Barbu	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	0.773	0.77
8	48	Argentina	0.707	0.714	0.719	0.725	0.729	0.731	0.738	0.746	0.752	0.763	0.77	0.775	0.77	0.775	0.775	0.777	0.80
9	81	Armenia	0.633	0.629	0.585	0.59	0.6	0.604	0.614	0.625	0.637	0.644	0.649	0.653	0.663	0.672	0.681	0.694	0.70
10	6	Australia	0.866	0.867	0.868	0.872	0.875	0.883	0.886	0.889	0.892	0.895	0.898	0.9	0.903	0.904	0.907	0.902	0.90
11	20	Austria	0.795	0.799	0.805	0.809	0.813	0.817	0.82	0.824	0.828	0.834	0.838	0.849	0.838	0.842	0.849	0.855	0.86
12	87	Azerbaijan	..	..	..	..	..	0.612	0.612	0.618	0.627	0.634	0.641	0.649	0.658	0.667	0.674	0.681	0.70
13	60	Bahamas	..	..	..	..	..	..	..	..	..	..	0.787	0.788	0.79	0.789	0.79	0.791	0.79
14	45	Bahrain	0.736	0.755	0.756	0.764	0.768	0.775	0.779	0.781	0.784	0.786	0.792	0.792	0.792	0.793	0.792	0.792	0.79
15	135	Bangladesh	0.388	0.395	0.403	0.411	0.419	0.427	0.436	0.444	0.453	0.462	0.47	0.479	0.485	0.492	0.499	0.506	0.51
16	56	Barbados	0.732	0.733	0.733	0.737	0.743	0.747	0.751	0.757	0.756	0.764	0.771	0.77	0.774	0.778	0.782	0.786	0.79
17	50	Belarus	..	..	..	..	..	0.656	0.661	0.667	0.671	0.676	0.682	0.689	0.696	0.704	0.714	0.724	0.7
18	17	Belgium	0.806	0.81	0.825	0.838	0.845	0.851	0.857	0.862	0.866	0.868	0.873	0.876	0.879	0.882	0.885	0.889	0.89
19	103	Belize	0.613	0.618	0.624	0.627	0.627	0.627	0.627	0.63	0.631	0.636	0.643	0.647	0.655	0.663	0.668	0.666	0.67
20	163	Benin	0.348	0.354	0.358	0.365	0.368	0.373	0.377	0.381	0.385	0.391	0.398	0.41	0.419	0.426	0.434	0.44	0.44
21	134	Bhutan	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	0.512	0.52
22	114	Bolivia (Plurinatic	0.54	0.549	0.555	0.562	0.57	0.578	0.585	0.587	0.599	0.608	0.616	0.619	0.625	0.629	0.63	0.632	0.63
23	75	Bosnia and Herzeg	..	..	..	..	..	..	..	..	..	..	0.669	0.675	0.681	0.686	0.692	0.7	0.70
24	94	Botswana	0.57	0.576	0.574	0.573	0.567	0.573	0.572	0.575	0.577	0.579	0.578	0.58	0.576	0.583	0.589	0.598	0.61
25	79	Brazil	0.613	0.62	0.626	0.634	0.642	0.651	0.657	0.664	0.67	0.675	0.684	0.691	0.698	0.694	0.697	0.7	0.70

Figure 1.3 Raw HDI dataset

- GFDDData.csv, which is the main dataset. This dataset is huge with lots of missing values and unnecessary columns and rows. A lot of effort has been put in order to clean this dataset.
- HDI.csv, this data set includes the information about the Human Development Index.
- GDP-POP.csv which contains information about the Gross Domestic Product which also includes the population for the respective years along with the population density.

M.2 Data Cleaning: The World Bank dataset required a lot of cleaning and filtering since it contained a lot of missing and irregular values. The cleaning method included removal of unnecessary columns and rows which had no importance in the dataset, insertion of an overall score metric which indicated how many missing values are present in a row. If the score is above a particular threshold, the row would be discarded because it is not possible to predict the value based on the erratic trend through the years.

```
project_data  <- read.csv("GFDDData.csv")
project_data  <- project_data [-c(5:49)]
project_data  <- subset(project_data , select = -c(X))
project_data ["Score"] <- rowSums(is.na(project_data ) | project_data  == "")

final_data <- filter(project_data , Score == 0)
colnames(final_data) <- c("Country Name", "Country Code", "Indicator Name", "Indicator Code", "2005"

table(final_data$`Country Name`)

final_data ["Sum_Value"] <- rowSums(final_data[,5:17])
final_data$Sum_Value=log(final_data$Sum_Value)
final_data <- filter(final_data , Sum_Value != '-Inf')
final_data <- filter(final_data , Sum_Value != 'NaN')

#gives frequency table of indicators per country
freqtable <- data.frame(table(final_data$`Country Name`))
colnames(freqtable) <- c("Country", "Freq")
```

Figure 1.4

Figure.1.4 is the code which was embedded to clear the missing data and removal of useless columns. The values are not number and infinite are filtered and given value indicators of each country. The missing values which were present had to be extracted. Some of the missing values which could not be removed had to be exchanged with those values which could be easily operated. Data cleaning was one of the most tedious and time-consuming process.

```
#hdi dataset
hdi <- read.csv("hdi.csv")
hdi <- hdi[-c(3:29)]
hdi <- hdi[-c(4)]
hdi <- hdi[-c(190:212),]
hdi <- data.frame(lapply(hdi, as.character), stringsAsFactors=FALSE)
hdi$Country[46] <- "Cote d'Ivoire"
colnames(hdi) <- c("HDI Rank", "Country", "HDI")
hdi <- hdi[-c(1)]

#gdp-population dataset
gdp <- read.csv("gdp-pop.csv")
gdp["Score"] <- rowSums(is.na(gdp) | gdp == "")
gdp <- filter(gdp, Score == 0)
gdp <- filter(gdp, Year == 2017)
gdp <- subset(gdp, select = -c(Score))
gdp <- data.frame(lapply(gdp, as.character), stringsAsFactors=FALSE)
colnames(gdp) <- c("Country", "Code", "Year", "Population Density", "GDP per capita", "Total Population")
gdp <- gdp[-c(2:4)]
```

Figure 1.5 HDI and GDP cleaning

The further cleaning of HDI and GDP dataset has been done by converting the factor datatype columns into strings. The useless and noisy data has been removed. These two datasets required less amount of work as compared to the GFDD.csv.

Clean Datasets -

	Country	GDP per capita	Total Population
1	Afghanistan	1803.98748708124	35530081
2	Albania	11803.4305936025	2930187
3	Algeria	13913.8393634819	41318142
4	Angola	5819.49497145261	29784193
5	Antigua and Barbuda	21490.9426586166	102012
6	Argentina	18933.9071474396	44271041
7	Armenia	8787.57993972036	2930450
8	Australia	44648.7099113362	24450561
9	Austria	45436.6858219914	8735453
10	Azerbaijan	15847.4188327529	9827589

	Country	HDI
1	Afghanistan	0.493
2	Albania	0.789
3	Algeria	0.758
4	Andorra	0.852
5	Angola	0.576
6	Antigua and Barbuda	0.774
7	Argentina	0.832

Interactive Visualization of World Bank – Global Financial Development Dataset

	Country	GDP per capita	Total Population	HDI	Freq
1	Afghanistan	1803.9875	35530081	0.493	19
2	Algeria	13913.8394	41318142	0.758	42
3	Angola	5819.4950	29784193	0.576	41
4	Antigua and Barbuda	21490.9427	102012	0.774	31
5	Argentina	18933.9071	44271041	0.832	71
6	Armenia	8787.5799	2930450	0.758	46
7	Australia	44648.7099	24450561	0.937	71
8	Austria	45436.6858	8735453	0.912	67
9	Azerbaijan	15847.4188	9827589	0.752	39
10	Bahrain	43290.7045	1492584	0.839	31
11	Bangladesh	3523.9839	164669751	0.609	43

M.3 Data Analysis: Data analysis is important to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis plays a very important role in order to organize the data so that our clean data is ready for the visualizing operations.

```
#merged dataframe for 1st layer
gh <- merge(gdp, hdi, by = "Country")
ghf <- merge(gh, freqtable, by = "Country")
ghf <- filter(ghf, Freq > 0)
ghf$`GDP per capita` <- as.numeric(ghf$`GDP per capita`)
ghf$`Total Population`<- as.numeric(ghf$`Total Population`)
ghf$HDI <- as.numeric(ghf$HDI)
colnames(ghf) <- c("Country", "GDP Per Capita", "Total Population", "HDI", "No. of Indicators")
Country_list <- ghf$Country
```

Figure 1.6 Merging of data frames

The unstructured clean data is structured. The values which are not number or infinite values are filtered. The final data set is the economic part. The dataset was provided with meaningful column names. The frequency table was created for the country column in the GFDD dataset so to check which countries had a substantial amount of indicators. The cleaned data-frames were merged along an INNER JOIN on the country column to find intersecting names. The final result was easy to interpret as the data had been structured and organized. It was possible to visualize data into meaningful information. The values that are not numbers and infinite are filtered and given value indicators of each country final data set is the economic part.

M.4 Data Visualization: The interactive data visualization had been performed by using R-shiny package in R-studio. Interactive data visualization is done in order to identify unnoticed information especially in large dataset like this one. If there is no visual data then the trends, behavior patterns and dependencies could be missed out. In this project an interactive Dash board has been created in order to visualize the trends and behavior patterns of the specific variables in the dataset. It is also possible to observe the dependencies that exist in the data and could be ignored with the visualization operations.

```
ui <- dashboardPage(
    dashboardHeader(title = "World Development Indicators", titleWidth = 300),
    dashboardSidebar(
        selectInput("country", "Select a country", choices= Country_list),
        selectInput("IndicatorValue", "Select an Indicator", "placeholder" ),
        collapsed = TRUE
    ),
```

Figure 1.7 Creating dashboard

M.4.1 Bubble Scatter Plot (1st layer):

The first layer includes the bubble scatter plot of the Human development index with respect to GDP per capita. The HDI was kept on the horizontal axis since it is an independent datatype whereas GDP per capita was put on vertical axis. The color saturation method has been used to reflect the varying values of HDI within its range from light shade of red to darker shade of red. The size range of bubble was kept from 10 to 50 in order to show the continuous variance in the population.

```
bubbleplot <- plot_ly(ghf, x = ~ghf$HDI, y = ~ghf$`GDP Per Capita`, type = 'scatter',
                      color = ~HDI, colors = 'Reds', mode = 'markers',
                      size = ~ghf$`Total Population`,sizes = c(10, 50),
                      marker = list(opacity = 0.5, sizemode = 'diameter'),
                      hoverinfo = 'text',
                      text = ~paste('Country:', ghf$Country,
                                    '<br>GDP per Capita :', ghf$`GDP Per Capita`,
                                    '<br>Population :', ghf$`Total Population`,
                                    '<br>HDI :', ghf$HDI,
                                    '<br>No. of Indicator :', ghf$`No. of Indicator`)
) %>% layout(xaxis = list(title = "Human Development Index"), yaxis = list(title = "GDP Per Capita"), legend = lis
print(bubbleplot)
```

Figure 1.8 Bubble Scatter Plot code

M.4.2 Bar-chart (2nd layer): The second layer is visualized through the bar-chart. The graph shows the summation of indicator values through the years against the indicator names. The Indicator names are put on the x-axis which is an independent variable. The y-axis has the dependent variable which has the sum values of the year’s corresponding to the country name. The country name is selected from the dash-board scroll bar options. The color option was chosen make the bar-chart look brighter and attractive.

```
barplot <- ggplot(filtered_data1, aes(y=filtered_data1$Sum_Value, x=filtered_data1$`Indicator Name`))+
    geom_bar(position="dodge", stat="identity", fill='#D81815')+
    coord_flip()+
    xlab("Indicator Names")+
    ylab("Sum of Values of Years")

barplot <- ggplotly(barplot)
print(barplot)
```

Figure 1.9 Bar chart code

M.4.3 Line Chart (3rd layer): The third layer is the line chart which is dynamic on both x and y axis. The line chart represents the plot about the values per year vs the time. The data plot varies as we change the country name and indicator of the particular country. The color of the line was chosen light blue with bright yellow point representing the data plots.

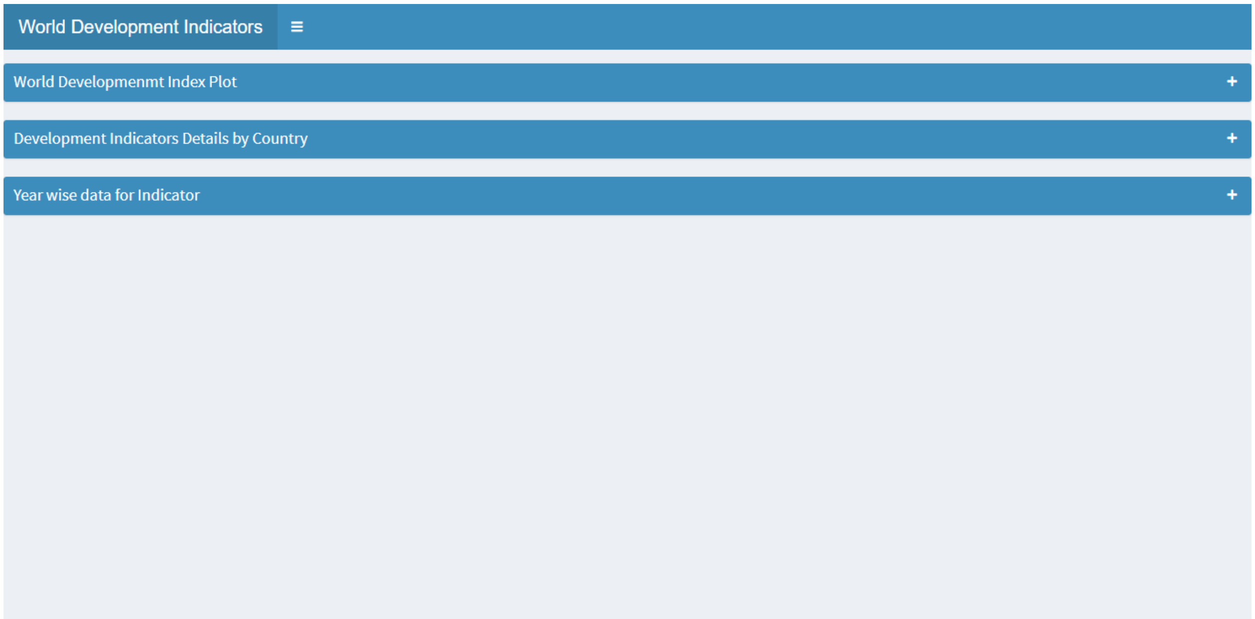
```
lineplot <- ggplot(modified_data, aes(x=period, y=obsValue, group = 1))+
    geom_point(color = '#FFC300', size = 5)+
    geom_line(color = '#0733B3', size = 0.8)+
    xlab("Years")+
    ylab("Values per Year")
print(lineplot)
```

Figure 1.10 Line Chart code

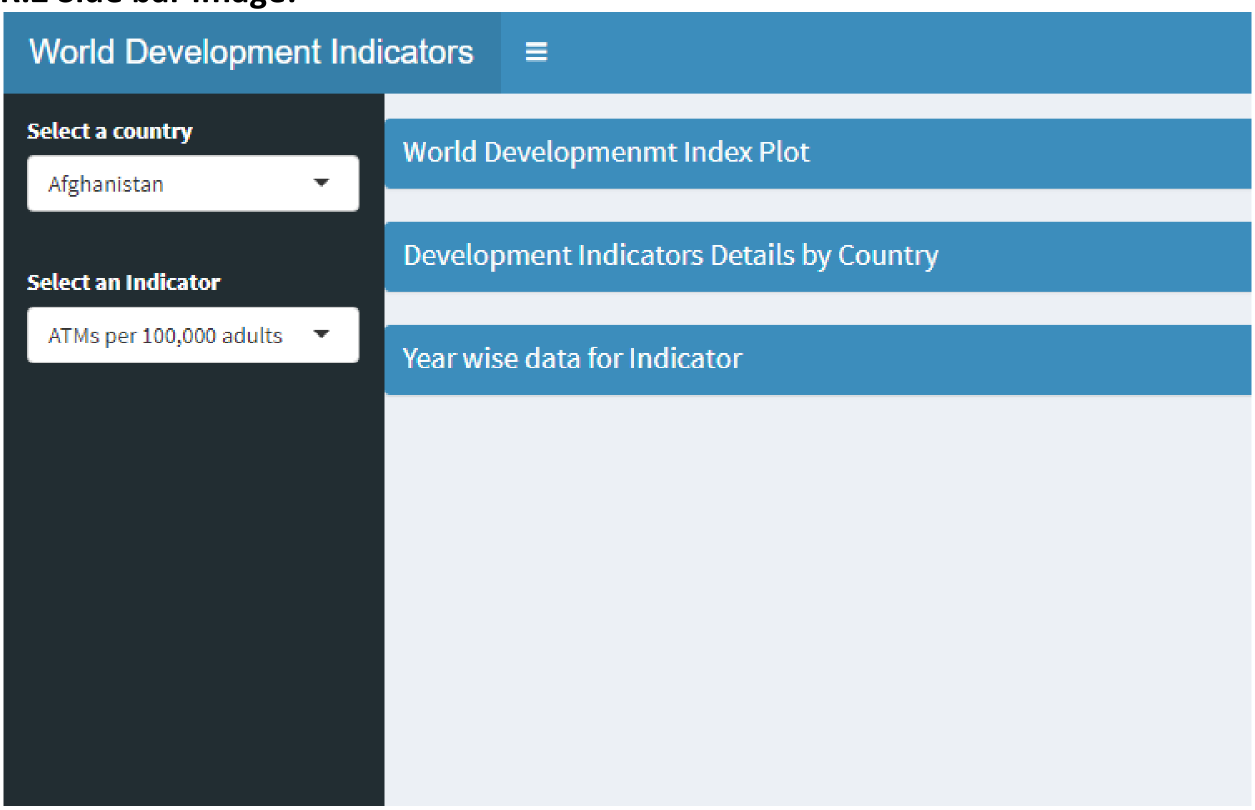
Results

R.1 Initial interactive dashboard image:

Interactive Visualization of World Bank – Global Financial Development Dataset

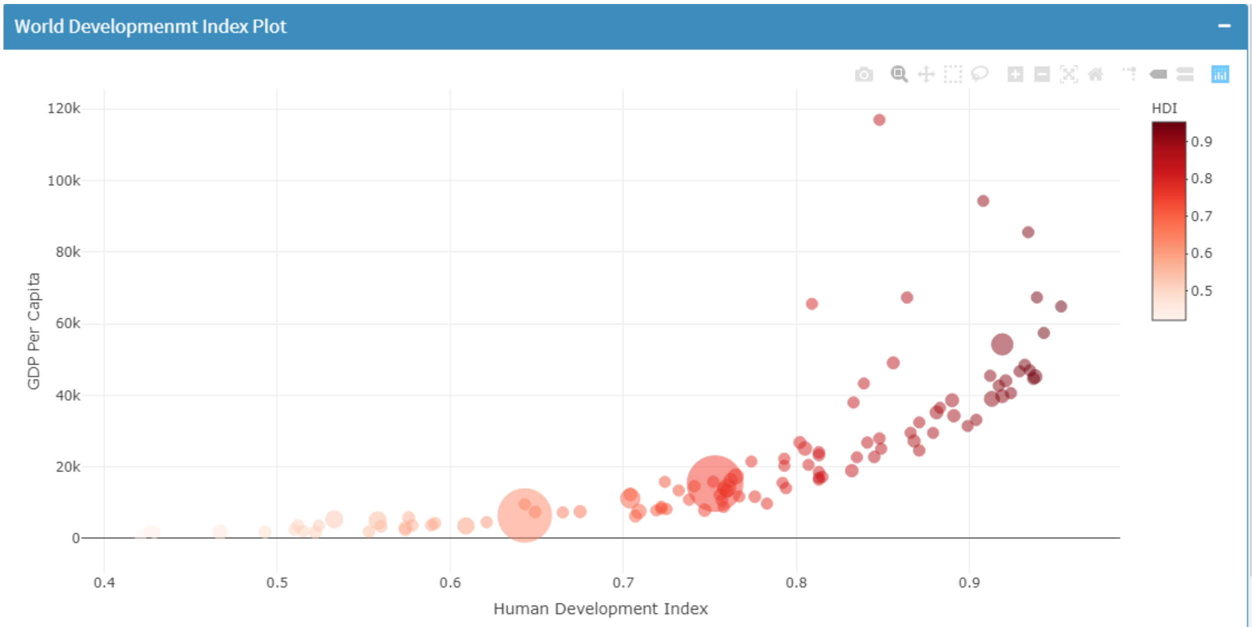


R.2 Side bar image:

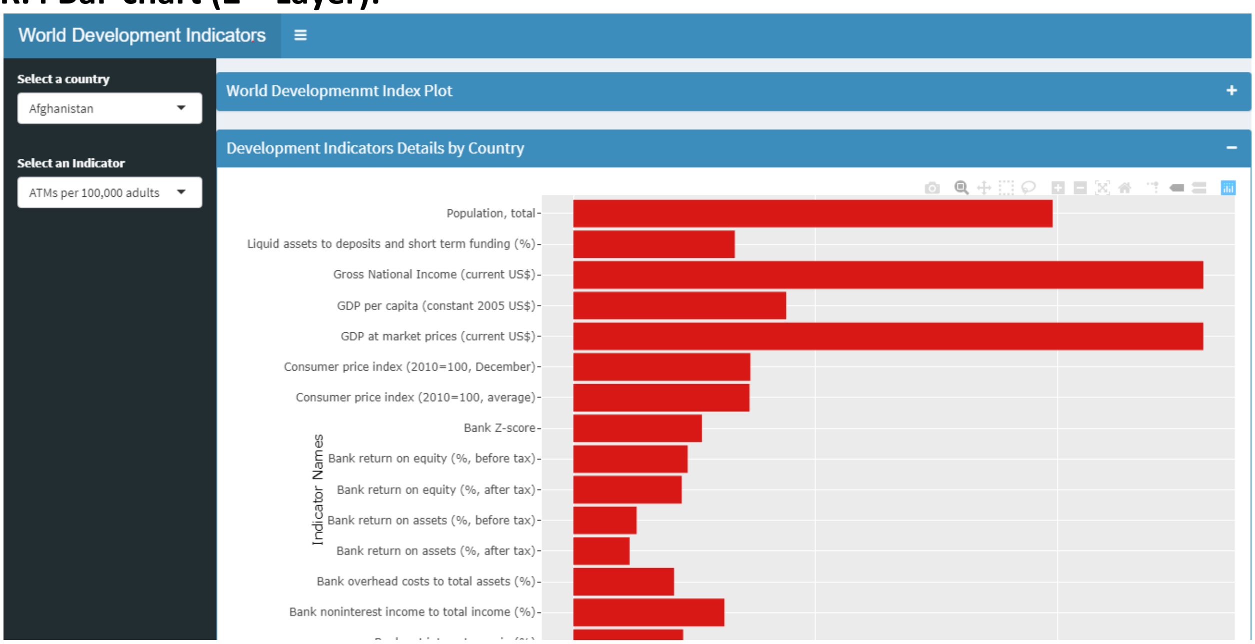


R.3 Bubble plot (1st Layer):

Interactive Visualization of World Bank – Global Financial Development Dataset

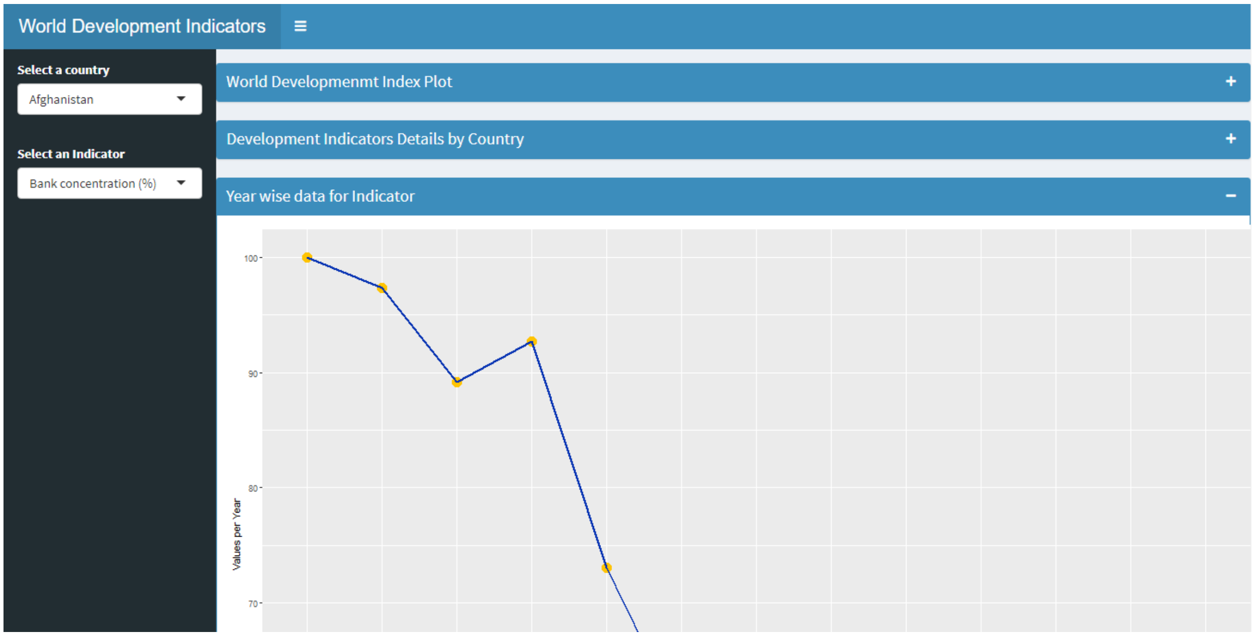


R.4 Bar-chart (2nd Layer):



R.5 Line chart(3rd Layer):

Interactive Visualization of World Bank – Global Financial Development Dataset



Division of Work:

Sushant Pagnis – Worked on data collection, data processing, cleaning and analysis. Developed first layer Bubble Plot using R, Shiny, Plotly.

Sanket Vora – Worked on data filtering, building Dashboard structure using Shiny, Shiny Dashboard. Developed second layer Bar plot using R, Shiny, ggplot, Plotly.

Swati Lathwal – Worked on developing third layer Line plot using R, Shiny, ggplot, Plotly. Prepared documentation for the project report and presentation.

Conclusion

Interactive Visualization of World Bank – Global Financial Development Dataset

Human beings are linear creatures we like to see how things develop and progress across time. Unfortunately, seeing data presented as a string of numbers does not generally allow for that kind of linearity. On the other hand, allowing users to interact with data presented in a clearly-visual manner, a data-intensive story becomes visible. Convenient and effective visualization is a necessity to harness the true power of data analysis. Interactive visualizations have a competitive advantage. In this study, Interactive data visualization allows users the freedom to fully explore the analyzed data, Users can manipulate the data to find out specific things they need to know, Users are presented with only the key elements that enable them to get both the big picture and the details in one visualization. Patterns make it easier for users to analyze the data and identify trends. It's unlikely that users would be able to recognize patterns when presented with millions of lines of data in a spreadsheet. Hence interactive visualizations make it easier to identify patterns at a glance.

References

https://datacatalog.worldbank.org/dataset/global-financial-development

http://hdr.undp.org/en/data#

https://ourworldindata.org/grapher/population-density-vs-prosperity