# TED TALKS: Individual Report

Submitted by: Anuj Verma

## Role:

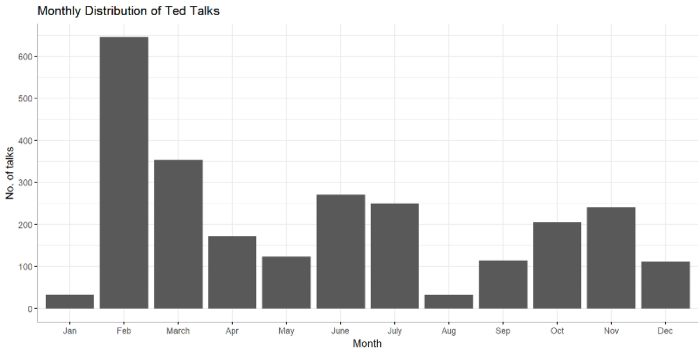
1. Analyze the trend in ted talks per month per year and see if any interesting patters emerge.
2. Analyze the most popular tags/topics in Ted conferences.

## Analysis:

### Part 1: Analyze the trend in number of ted talks

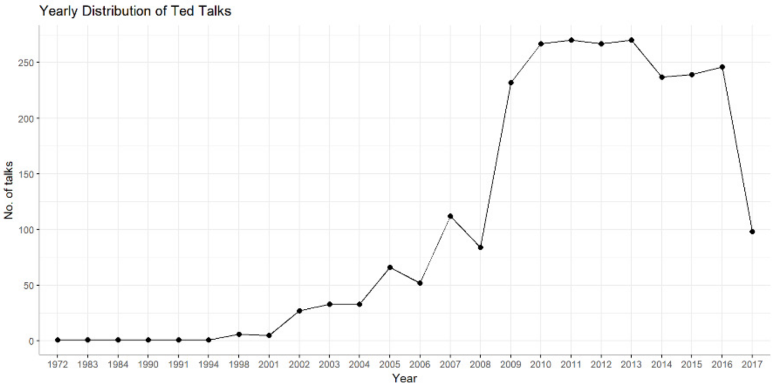
First, I created the monthly distribution of ted talks using bar graph in R. The data is categorized based on months only irrespective of the year number.

**Interpretation:** The bar graph shows overall distribution of the number of talks during the months. We can see that February month is very famous for delivering ted talks. This is because the Ted official conference is held in this month. January and August are the least popular months for delivering talks.

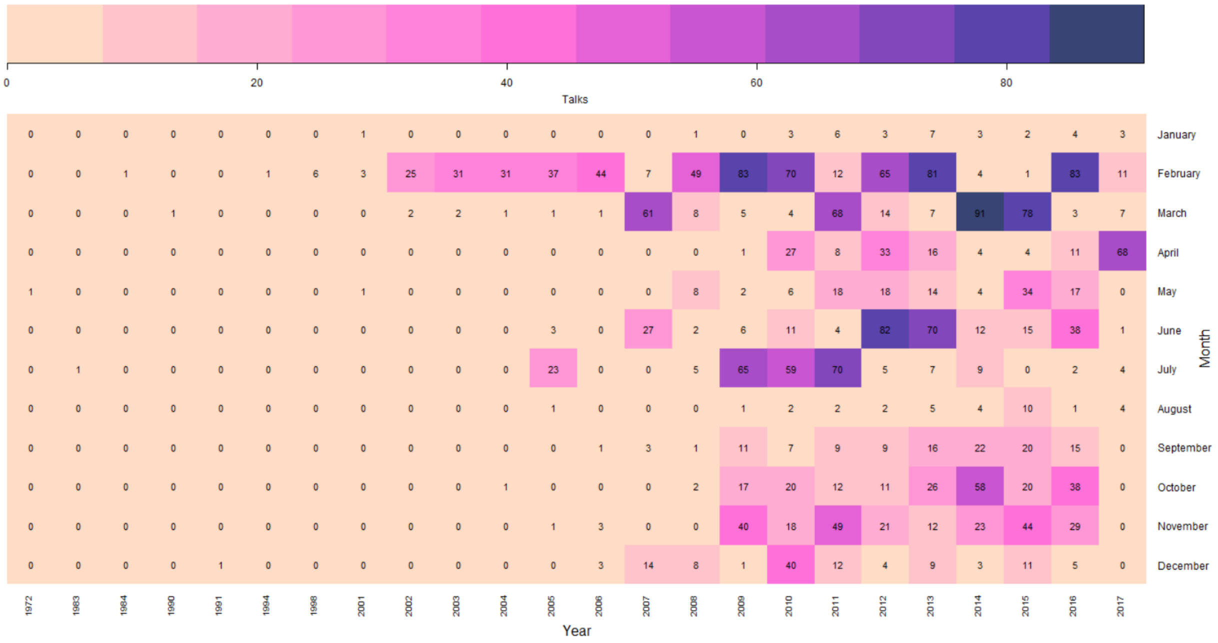


I also wanted to see the yearly distribution of number of talks and decided to explore it using line-point graph in R.

**Interpretation:** The graph explores the number of talks delivered each year. The number of talks increased gradually from 1972-2008, but then the number of talks delivered in 2009 increased by more than 2 folds. This could be because of sudden increase in popularity of Ted talks. The number of talks have been similar after that. The sudden decrease in 2017 could be because of insufficient data.



Next, I wanted to analyze whether number of talks per Month each year shows any other patterns that may not be visible in the above individual graphs. So, I decided to plot a heatmap to map months, years and number of talks on one map. I used heatmap.2 function from library(gplots) in R.

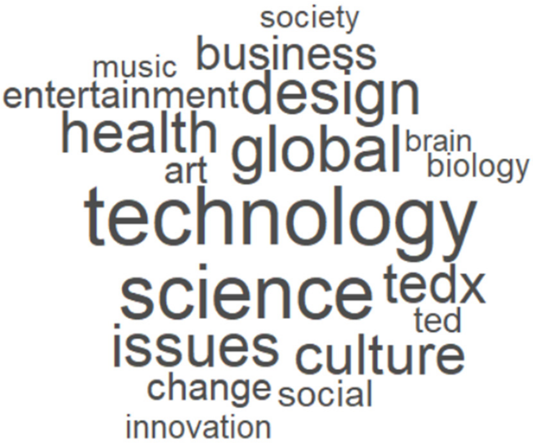


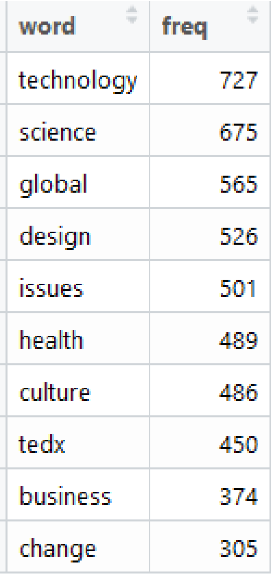
**Interpretation:** The above heatmap displays the summary of talks for months and years. The color gradient is described by the number of talks. The dark purple means more talks and shading purple/sand color denotes less number of talks as clearly labeled in the legend. We can see that February is the most popular month for talks with high number of talks per year while January and August are least popular months with least number of talks per year. High talks in February are because the official ted conferences are held in this month. The high number of talks for few other months are because of the world-wide ted conferences. We see more purple color towards right of the graph since number of talks increased in these years.

## Part 2: Analyze the Most Popular Topics in Ted Talks

I wanted to visualize the most popular topics using a word cloud. The challenge in finding most popular topic was that there are multiple tags associated with single presentation, which is what generally happens. The word cloud was created with using the text parser on ‘Tags’ column. The text parser created a text file of all the tags and a frequency table for the tags was created from this text file. However, in this process, there are some compromises that I had to make. The tags with 2 words were split into two 1 word tags. For example, a tag named ‘Global Warming’ was split into two separate tags ‘Global’ and ‘Warming’. This is a work for future.

After processing the data below word cloud was generated.

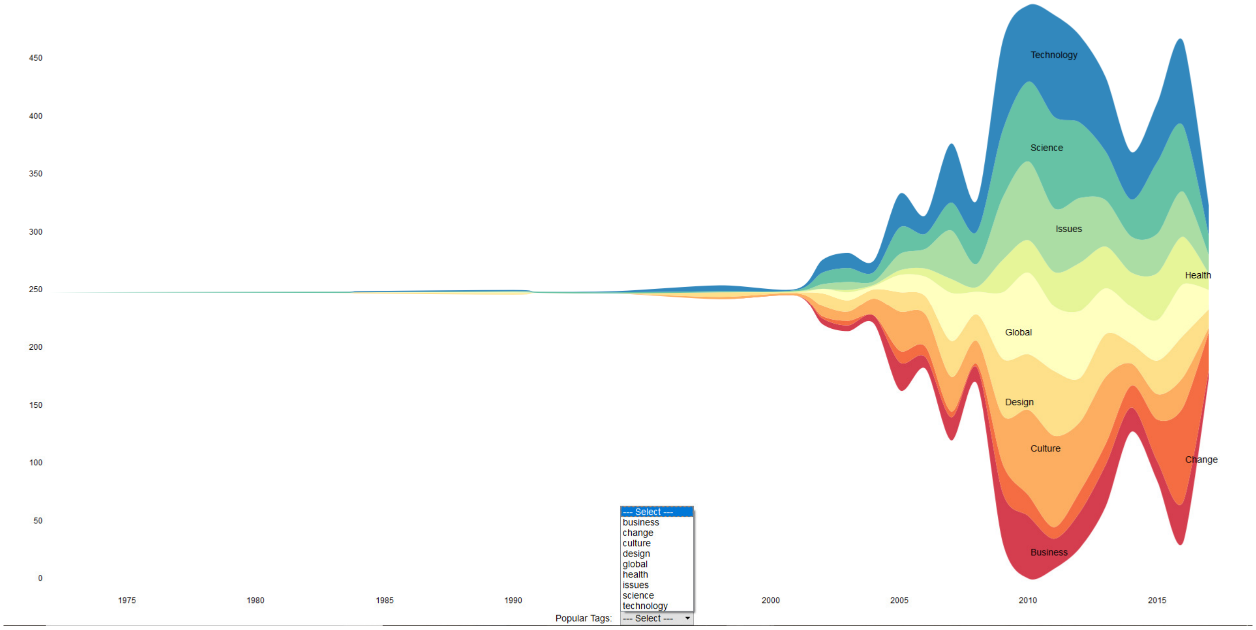




| word | freq |
|---|---|
| technology | 727 |
| science | 675 |
| global | 565 |
| design | 526 |
| issues | 501 |
| health | 489 |
| culture | 486 |
| tedx | 450 |
| business | 374 |
| change | 305 |

**Interpretation:** The above word cloud shows the popular tags in Ted conferences. This cloud contains only those tags which were used more than 200 times during the Ted history. The font is bigger for higher frequency of use and smaller for lower frequency of use. The word cloud analysis was taken one step further by generating the dataset for most frequent topics.  Only top 10 tags were picked from this and Tedx tags were removed since they represent the general name of event and not the topics.

Next question I wanted to answer was how these popular topics were used over the period of Ted talk history. To answer this, I chose Stream graph since it is an interactive tool which can display the popularity of these topics over the years like no other graph can does.

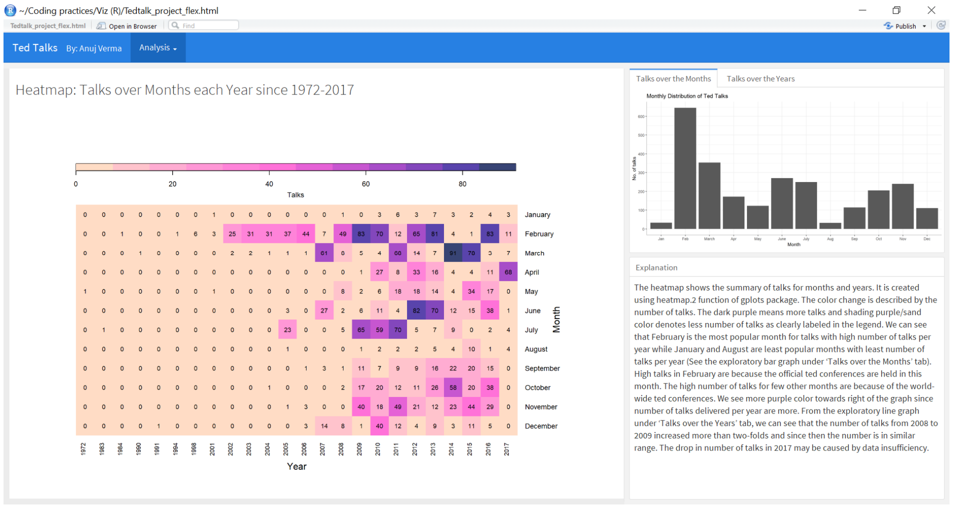


**Interpretation:** The above stream graph visualization depicts the use of tags/topics over the years (1972-2017). It is made using the streamgraph html widget package available at github. This is an interactive graph and user can hover mouse on the parts of the graph to see the number of talks in the corresponding year related to the topic. User can also select a tag from the Popular tags drop down; this will highlight the selected tag's color band in the graph. The stream graph only displays the popularity of only Top 9 topics. Each topic is represented by different color and is labeled accordingly. The width of the color band represents the number of talks on that topic in corresponding year. A sharp increase in use of these topics can be seen at 2009 (as analysed in previous heatmap). This is simply because the total number of ted conferences increased during this period. The graph suggests that the most number of talks were from 2009-2013 since the peak is highest for these years. The topics Technology, Science, Global, Design, Culture and Business were discussed most. One interesting topic to observe is 'Change'. It was not much popular throughout the Ted history, but it was certainly discussed most in 2016. The nature of change is not clear, and further analysis can be performed to know whether it is related to climate change, society change or some other change.

## Flexdashboard

I decided to create a flexdashboard to display the visualizations in a better user-friendly structure. My dashboard will have two pages under 'Analysis' menu bar. First page shows the number of ted-talks analysis and second page shows the interactive stream graph of popular topics.

Page-1: Number of Talks



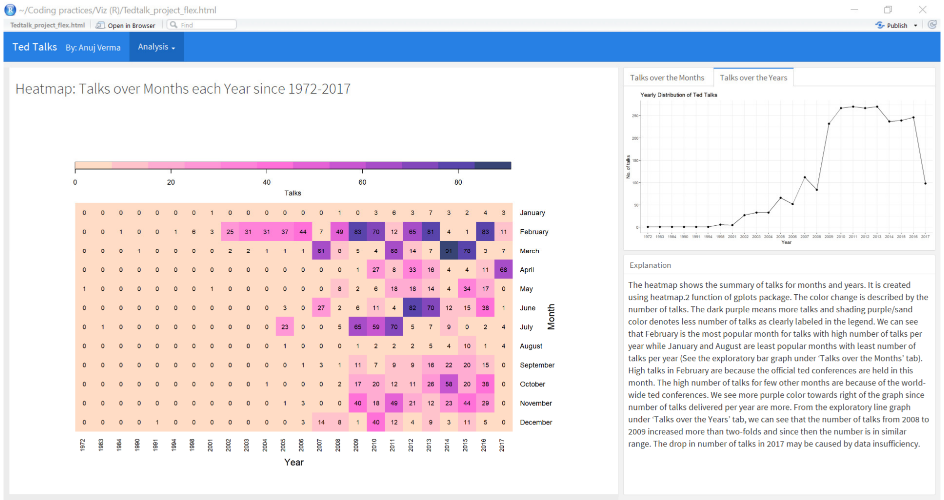
Heatmap: Talks over Months each Year since 1972-2017

| Month | 1972 | 1983 | 1984 | 1990 | 1991 | 1994 | 1998 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| January | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 6 | 3 | 7 | 3 | 2 | 4 | 3 |
| February | 0 | 0 | 1 | 0 | 0 | 1 | 6 | 3 | 25 | 31 | 31 | 37 | 44 | 7 | 49 | 83 | 70 | 12 | 65 | 81 | 4 | 1 | 83 | 11 |
| March | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 1 | 61 | 8 | 5 | 4 | 68 | 14 | 7 | 91 | 78 | 3 | 7 |
| April | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 27 | 8 | 33 | 16 | 4 | 4 | 11 | 68 |
| May | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 2 | 6 | 18 | 18 | 14 | 4 | 34 | 17 | 0 |
| June | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 27 | 2 | 6 | 11 | 4 | 82 | 70 | 12 | 15 | 38 | 1 |
| July | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 5 | 65 | 59 | 70 | 5 | 7 | 9 | 0 | 2 | 4 |
| August | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 2 | 2 | 5 | 4 | 10 | 1 | 4 |
| September | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 11 | 7 | 9 | 9 | 16 | 22 | 20 | 15 | 0 |
| October | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 17 | 20 | 12 | 11 | 26 | 58 | 20 | 38 | 0 |
| November | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 40 | 18 | 49 | 21 | 12 | 23 | 44 | 29 | 0 |
| December | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 14 | 8 | 1 | 40 | 12 | 4 | 9 | 3 | 11 | 5 | 0 |

Year

Explanation

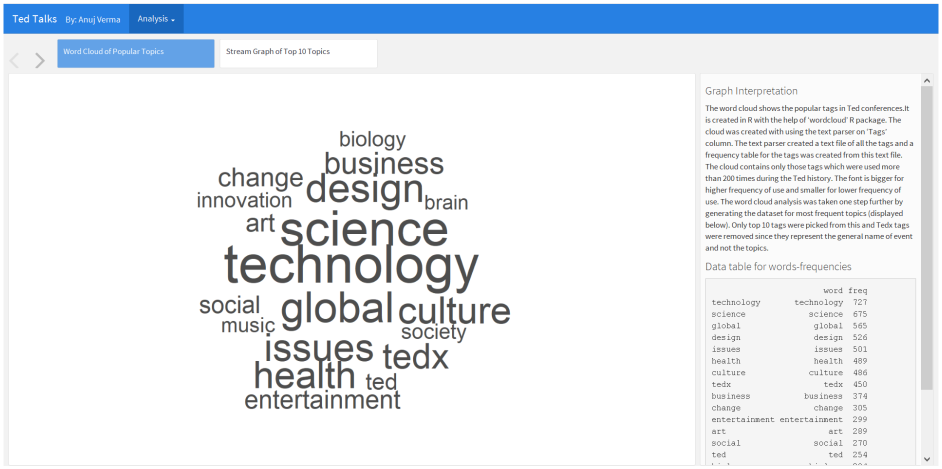
The heatmap shows the summary of talks for months and years. It is created using heatmap.2 function of gplots package. The color change is described by the number of talks. The dark purple means more talks and shading purple/sand color denotes less number of talks as clearly labeled in the legend. We can see that February is the most popular month for talks with high number of talks per year while January and August are least popular months with least number of talks per year (See the exploratory bar graph under 'Talks over the Months' tab). High talks in February are because the official ted conferences are held in this month. The high number of talks for few other months are because of the world-wide ted conferences. We see more purple color towards right of the graph since number of talks delivered per year are more. From the exploratory line graph under 'Talks over the Years' tab, we can see that the number of talks from 2008 to 2009 increased more than two-folds and since then the number is in similar range. The drop in number of talks in 2017 may be caused by data insufficiency.


~/Coding practices/Viz (R)/Tedtalk_project_flex.html
Tedtalk_project_flex.html
Open in Browser
Find
Publish
Ted Talks
By: Anuj Verma
Analysis
Heatmap: Talks over Months each Year since 1972-2017
0
20
40
60
80
Talks
0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 3 6 3 7 3 2 4 3 January
0 0 1 0 0 1 6 3 25 31 31 37 44 7 49 83 70 12 65 81 4 1 83 11 February
0 0 0 1 0 0 0 0 2 2 1 1 1 61 8 5 4 68 14 7 91 78 3 7 March
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 27 8 33 16 4 4 11 68 April
1 0 0 0 0 0 0 1 0 0 0 0 0 0 8 2 6 18 18 14 4 34 17 0 May
0 0 0 0 0 0 0 0 0 0 0 3 0 27 2 6 11 4 82 70 12 15 38 1 June
0 1 0 0 0 0 0 0 0 0 0 23 0 0 5 65 59 70 5 7 9 0 2 4 July
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 2 2 2 5 4 10 1 4 August
0 0 0 0 0 0 0 0 0 0 0 0 1 3 1 11 7 9 9 16 22 20 15 0 September
0 0 0 0 0 0 0 0 0 0 1 0 0 0 2 17 20 12 11 26 58 20 38 0 October
0 0 0 0 0 0 0 0 0 0 0 1 3 0 0 40 18 49 21 12 23 44 29 0 November
0 0 0 0 1 0 0 0 0 0 0 0 3 14 8 1 40 12 4 9 3 11 5 0 December
Month
1972 1983 1984 1990 1991 1994 1998 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year
Talks over the Months
Talks over the Years
Yearly Distribution of Ted Talks
No. of talks
0
50
100
150
200
250
1972 1983 1984 1990 1991 1994 1998 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year

Explanation

The heatmap shows the summary of talks for months and years. It is created using heatmap.2 function of gplots package. The color change is described by the number of talks. The dark purple means more talks and shading purple/sand color denotes less number of talks as clearly labeled in the legend. We can see that February is the most popular month for talks with high number of talks per year while January and August are least popular months with least number of talks per year (See the exploratory bar graph under 'Talks over the Months' tab). High talks in February are because the official ted conferences are held in this month. The high number of talks for few other months are because of the world-wide ted conferences. We see more purple color towards right of the graph since number of talks delivered per year are more. From the exploratory line graph under 'Talks over the Years' tab, we can see that the number of talks from 2008 to 2009 increased more than two-folds and since then the number is in similar range. The drop in number of talks in 2017 may be caused by data insufficiency.

## Page-2: Popular Topics

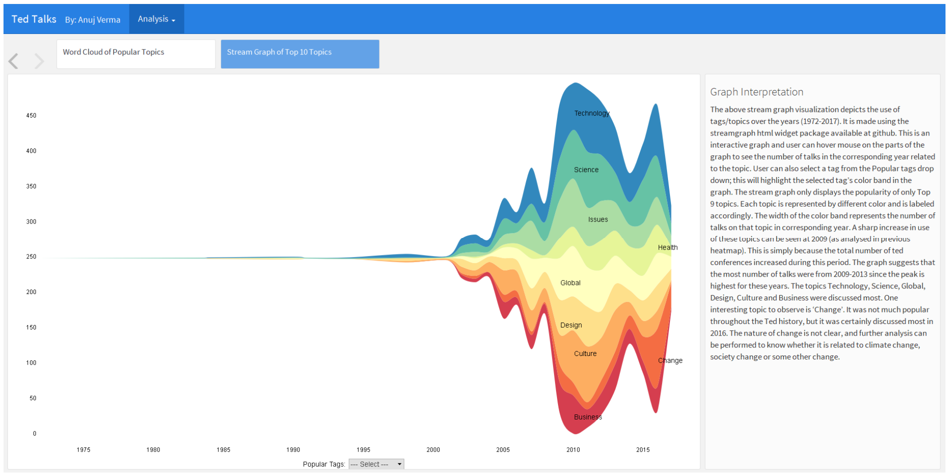


Graph Interpretation

The word cloud shows the popular tags in Ted conferences.It is created in R with the help of 'wordcloud' R package. The cloud was created with using the text parser on 'Tags' column. The text parser created a text file of all the tags and a frequency table for the tags was created from this text file. The cloud contains only those tags which were used more than 200 times during the Ted history. The font is bigger for higher frequency of use and smaller for lower frequency of use. The word cloud analysis was taken one step further by generating the dataset for most frequent topics (displayed below). Only top 10 tags were picked from this and Tedx tags were removed since they represent the general name of event and not the topics.

Data table for words-frequencies

```
                                word freq
technology       technology  727
science             science  675
global               global  565
design               design  526
issues               issues  501
health               health  489
culture             culture  486
tedx                   tedx  450
business           business  374
change               change  305
entertainment entertainment  299
art                     art  289
social               social  270
ted                     ted  254
```



Graph Interpretation

The above stream graph visualization depicts the use of tags/topics over the years (1972-2017). It is made using the streamgraph html widget package available at github. This is an interactive graph and user can hover mouse on the parts of the graph to see the number of talks in the corresponding year related to the topic. User can also select a tag from the Popular tags drop down; this will highlight the selected tag's color band in the graph. The stream graph only displays the popularity of only Top 9 topics. Each topic is represented by different color and is labeled accordingly. The width of the color band represents the number of talks on that topic in corresponding year. A sharp increase in use of these topics can be seen at 2009 (as analysed in previous heatmap). This is simply because the total number of ted conferences increased during this period. The graph suggests that the most number of talks were from 2009-2013 since the peak is highest for these years. The topics Technology, Science, Global, Design, Culture and Business were discussed most. One interesting topic to observe is 'Change'. It was not much popular throughout the Ted history, but it was certainly discussed most in 2016. The nature of change is not clear, and further analysis can be performed to know whether it is related to climate change, society change or some other change.

## Conclusion:

Even though TED initially started as ideas sharing platform for Technology, Entertainment and Design fields, it has not been limited to only these fields. When Ted started becoming popular more and more presenters from other fields also started to share their ideas/view. We can see that topics like science, culture, business, issues, health, change, global etc. are equally popular among the presenters.

## Learnings:

I learned the importance of a good visualization, enabling me to think critically while making graphs with respect to the audience so that the idea is communicated easily. During this project I learned how to create word cloud which was always intriguing to me with the bonus of stream graph which I feel is such a great interactive tool to show the change over a period of time. I learned about many R packages that makes our life simple and creates advanced graphs with such ease. There are still so many graph techniques that I could have tried on this dataset. I wanted to build a network plot but, due to time restrictions, I am not able to build a network structure at this point.

## References

Heatmap.2:

https://www.rdocumentation.org/packages/gplots/versions/3.0.1/topics/heatmap.2

https://stackoverflow.com/questions/24621070/heatmap-2-with-color-key-on-top

http://www.molecularecologist.com/2013/08/making-heatmaps-with-r-for-microbiome-analysis/

WordCloud:

http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

https://www.r-bloggers.com/building-wordclouds-in-r/

https://stats.stackexchange.com/questions/164372/what-is-vectorsource-and-vcorpus-in-tm-text-mining-package-in-r

StreamGraph:

https://stackoverflow.com/questions/13084998/streamgraphs-in-r

https://hrbrmstr.github.io/streamgraph/

https://github.com/hrbrmstr/streamgraph

https://rud.is/b/2015/03/12/streamgraph-htmlwidget-version-0-7-released-adds-support-for-markers-annotations/

https://www.rdocumentation.org/packages/tidyr/versions/0.8.0/topics/gather

Flexdashboard:

https://rmarkdown.rstudio.com/flexdashboard/layouts.html

https://stackoverflow.com/questions/36451484/how-to-combine-row-and-column-layout-in-flexdashboard?noredirect=1&lq=1

Other References:

https://www.kaggle.com/rounakbanik/ted-data-analysis

https://sebastiansauer.github.io/figure_sizing_knitr/