Data Visualization: Tale of data through visuals

India Autumn SDE,
Saturday, 18th Sept 2021

Supriya Moghe

accenture

Data Visualization: Tale of data through visuals

Agenda

This paper covers the story of data and emphasizes the importance of visualization techniques.

01 Abstract

02 The story of ‘Data’

03 Warm-up

04 Why do we need visualization techniques?

05 Some graphs (Waterfall Chart, bar chart, scattered plot)

06 Conference Success - Illustration

07 Interactive Chart

08 References

Abstract | Data visualization – a tale of data through visuals

Every year more than 200K trials are initiated. Clinical trial data is increasing exponentially. Well, when one says exponentially it means, if the data was plotted, it would reveal a certain pattern.

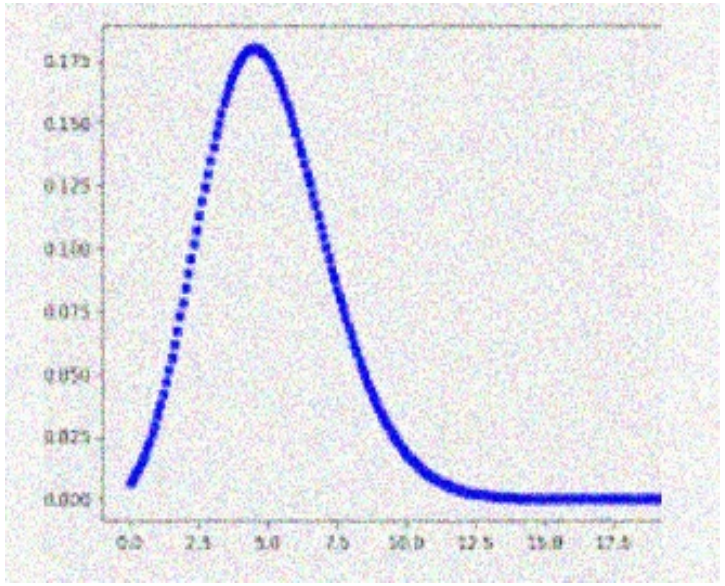
Data Visualization or **Infographics** is a technique of interpreting the data with the help of information from the graphs. It helps to check data models, outliers and spread of the data. This enables us to determine the statistical techniques which could be used for analysis. Advanced tools, like **R**, **Python** and **Tableau**, enable users to view the data in various ways, dissect it by several parameters in consideration, with just a click. Visuals such as automated graphs and interactive dashboards empower users to monitor and aggregate patient data from the constantly changing enormous data and understand the "**Big picture**".

This paper will focus on how data readability varies with change in visual representation and how it aids in inference. Let us explore the power of visuals.

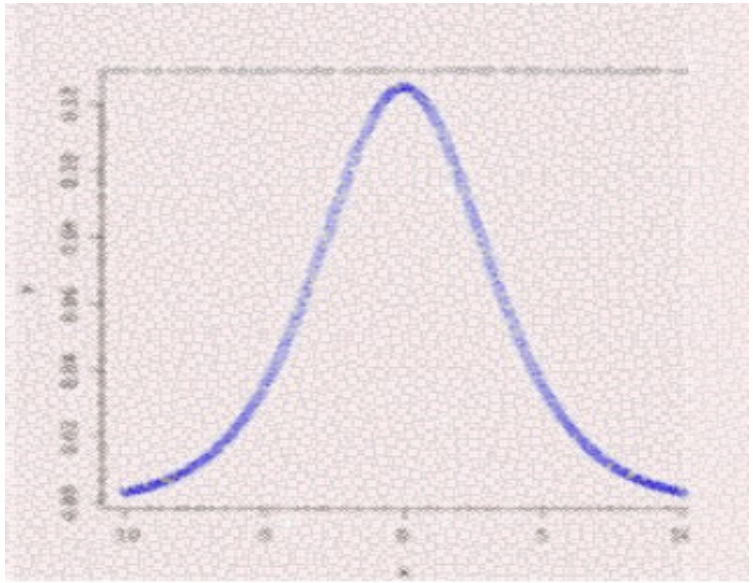
The story of ‘Data’

Warm up quiz

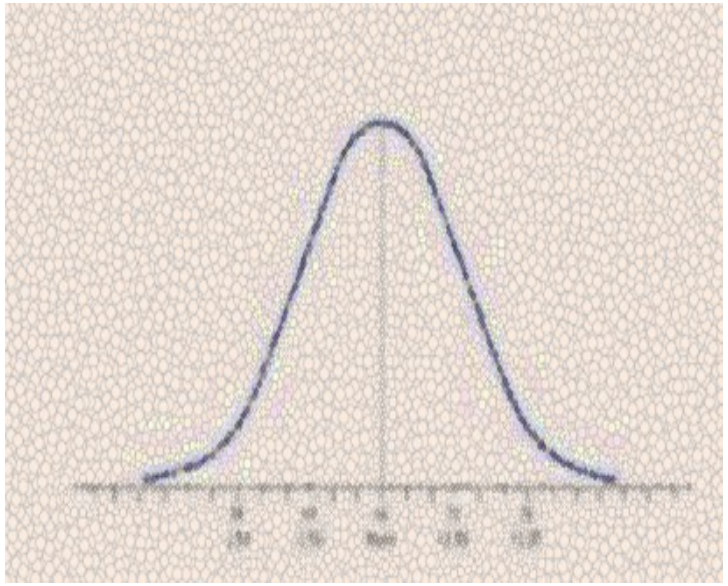
Identify the distribution in the graphs below.



Poisson



Logistic regression



Normal

Graph helped us to identify the distribution of the data.

Why do we need visualization techniques?

Transform row data into summary leading to centralized tendency of data

Cumulative impact

Support scientific claim

Derive more value from large amounts of data

Identify relationships, patterns or emerging trends and outliers too

Simplify to infer leading to **O**ptimized decision making

Bigger picture when multiple datapoints are presented together **OR** Bigger picture when the same datapoints are presented for long time-periods

Better Presentation

A Picture Is Worth a Thousand Words

Waterfall plot

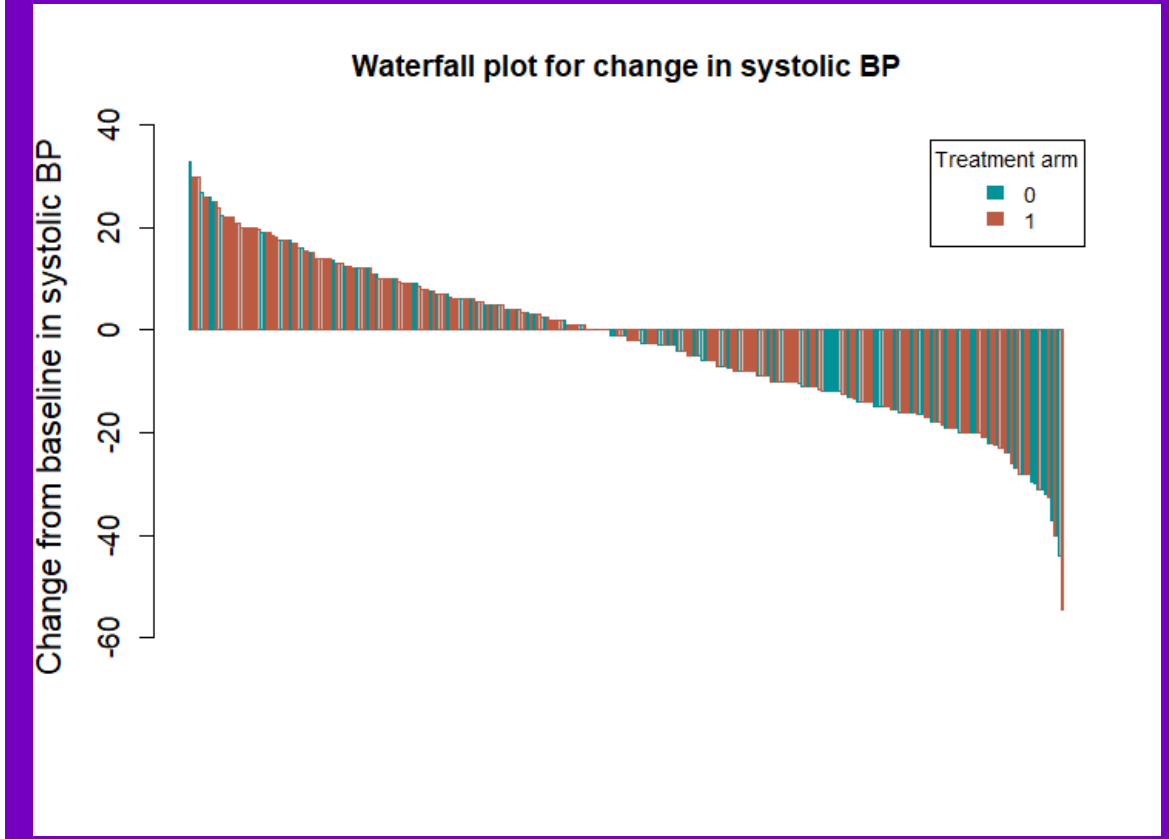
Code

```r
#set library for plot
library(ggplot2)
library(dplyr)
library(gapminder)

#Convert Male into facotr varible
heart_2$male <- as.factor(heart_2$male)

#Order response variable in descending order
o <- order(heart_2$sysbp12,decreasing=TRUE,na.last=NA)
heart_2 <- heart_2[o,]
col <- ifelse(heart$male == 0, "#BC5A42", "#009296")

barplot(heart_2$sysbp12, border=col, space=0.5, ylim=c(-50,30),
        main = "Waterfall plot for change in systolic BP", ylab="Change from baseline in systolic BP",
        cex.axis=1.2, cex.lab=1.4, legend.text=c(0,1),
        args.legend=list(title="Treatment arm", fill=col, border=NA, cex=0.9))
```

Sample Data

male	age	diabetes	totChol	sysbp11	sysbp12	diaBP	BMI	heartRate	glucose	TenYearCHD	ctr	yr_cat
1	39	0	195	106	14	70	26.97	80	77	0	China	1990
0	46	0	250	121	15	81	28.73	95	76	0	Thailand	1990
1	48	0	245	127.5	-7.5	80	25.34	75	70	0	Japan	1990
0	61	0	225	150	-20	95	28.58	65	103	1	South Cor	1990
0	46	0	285	130	10	84	23.1	85	85	0	China	1991
0	43	0	228	180	-20	110	30.3	77	99	0	Thailand	1991
0	63	0	205	138	18	71	33.11	60	85	1	Japan	1991
0	45	0	313	100	-20	71	21.68	79	78	0	South Cor	1991
1	52	0	260	141.5	19	89	26.36	76	79	0	Australia	1991
1	43	0	225	162	-40	107	23.61	93	88	0	United Sta	1991
0	50	0	254	133	-13	76	22.91	75	76	0	Singapore	1991
0	43	0	247	131	-11	88	27.64	72	61	0	Malaysia	1991

Output



Bar racing chart

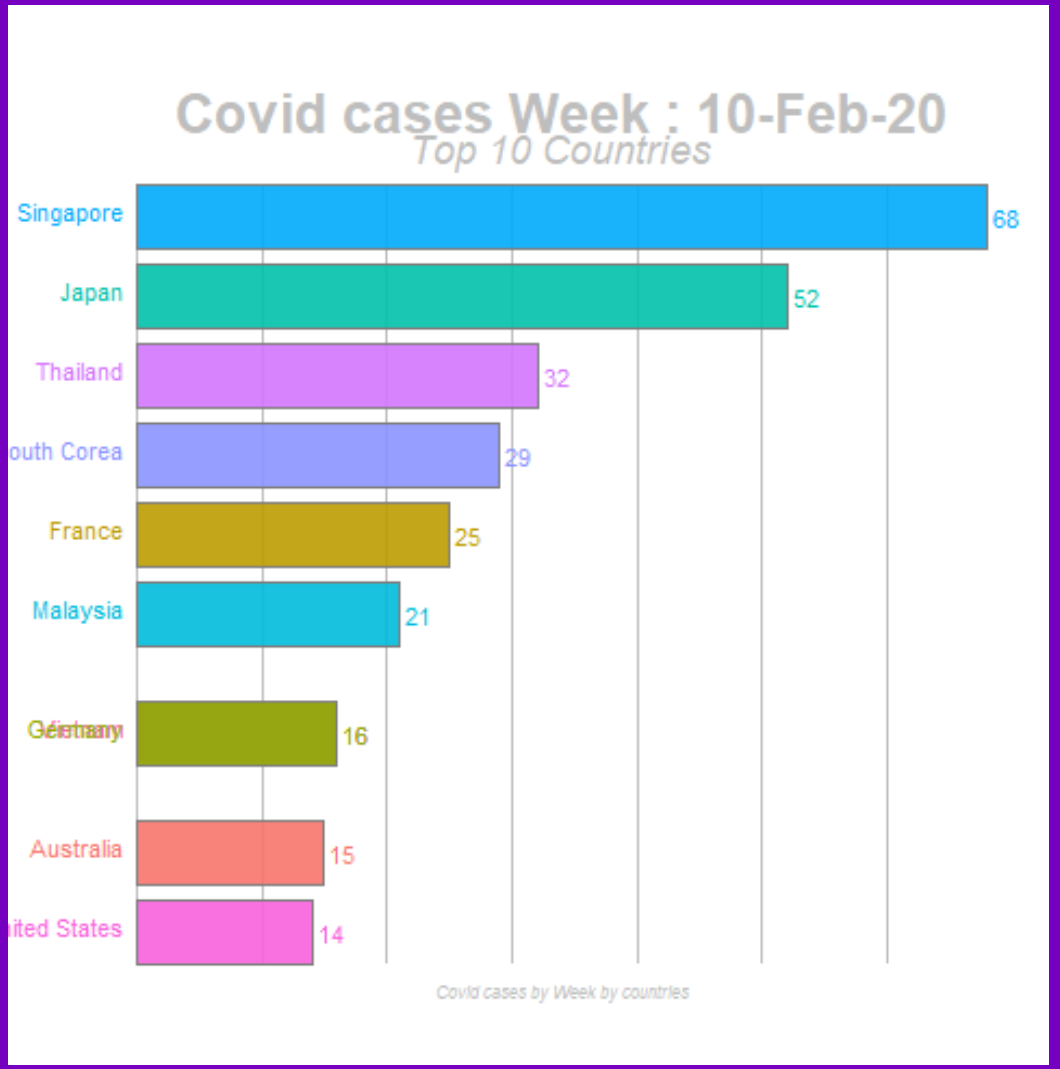
Code

```
R 4.1.0 · ~/
> covid_2 <- covid %>%
+   group_by(covid$week) %>% # The * 1 makes it possible to have non-integer ranks while sliding
+   mutate(rank = rank(-cases),
+          Value_rel = cases/cases[rank==1],
+          Value_lbl = paste0(" ",round(cases/1e9))) %>%
+   group_by(countryname) %>%
+   filter(rank <=5) %>%
+   ungroup()
>
> staticplot = ggplot(data=covid_2)+
+   aes(covid_2$cases , rank, group = countryname, fill = as.factor(countryname), color = as.factor(countryname)) +
+   geom_tile(aes(y = cases/2,
+                 height = cases,
+                 width = 0.9), alpha = 0.8, color = NA) +
+   geom_text(aes(y = 0, label = paste(countryname, " ")), vjust = 0.2, hjust = 1) +
+   geom_text(aes(y=cases, label = Value_lbl, hjust=0)) +
+
+   coord_flip(clip = "off", expand = FALSE) +
+
+   scale_y_continuous(labels = scales::comma) +
+   scale_x_reverse() +
+
+   theme(axis.line=element_blank(),
+         axis.text.x=element_blank(),
+         axis.text.y=element_blank(),
+         axis.ticks=element_blank(),
+         axis.title.x=element_blank(),
+         axis.title.y=element_blank(),
+         legend.position="none",
+         panel.background=element_blank(),
+         panel.border=element_blank(),
+         panel.grid.major=element_blank(),
+         panel.grid.minor=element_blank(),
+         panel.grid.major.x = element_line( size=.1, color="grey" ),
+         panel.grid.minor.x = element_line( size=.1, color="grey" ),
+         plot.title=element_text(size=25, hjust=0.5, face="bold", colour="grey", vjust=-1),
+         plot.subtitle=element_text(size=18, hjust=0.5, face="italic", color="grey"),
+         plot.caption =element_text(size=8, hjust=0.5, face="italic", color="grey"),
+         plot.background=element_blank(),
+         plot.margin = margin(2,2, 2, 4, "cm"))
>
>
> animate(anim, 200, fps = 20,  width = 1200, height = 1000,
+         renderer = gifski_renderer("gganim.gif"))
```

Sample Data

countryname	week	cases
Thailand	21-Jan-20	2
Japan	21-Jan-20	1
South Corea	21-Jan-20	1
Thailand	27-Jan-20	14
Japan	27-Jan-20	15
South Corea	27-Jan-20	11
Australia	27-Jan-20	19

Output



Animated scattered plot

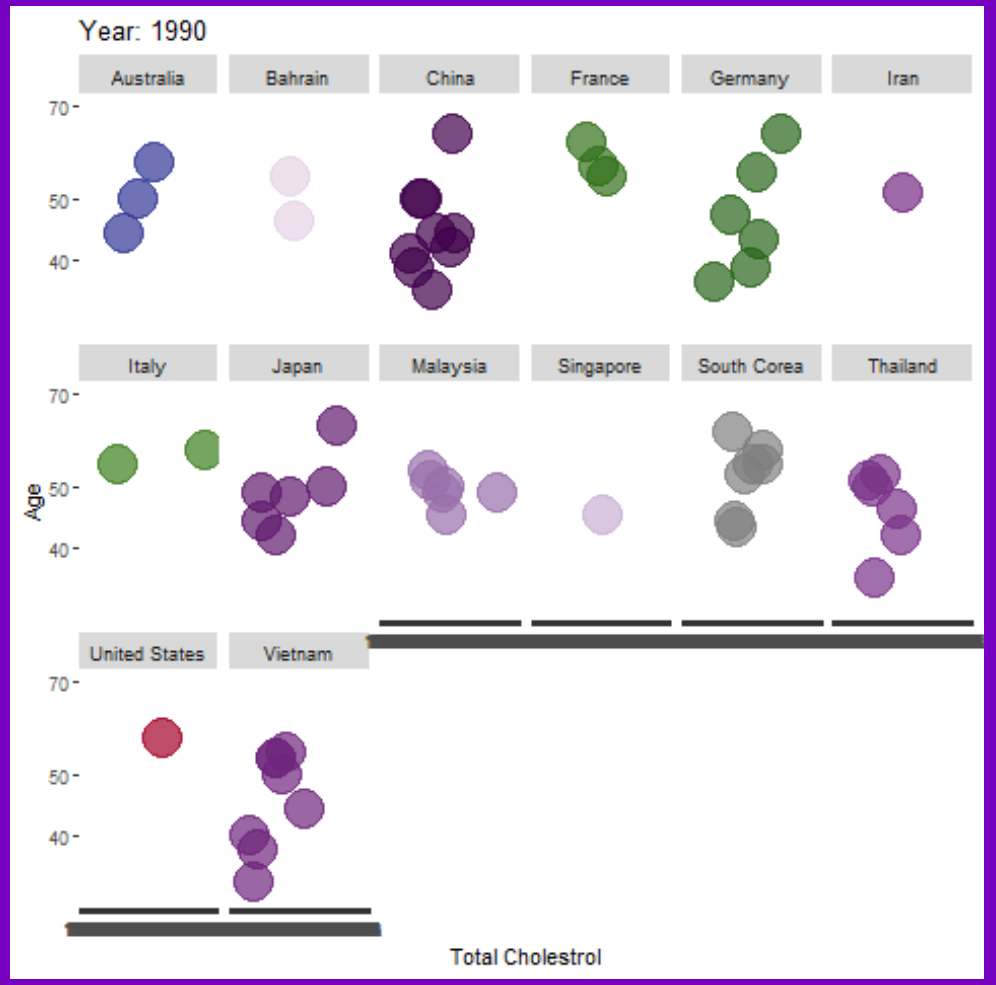
Code

```
##Import below libraries which are needed to plot the graphs
install.packages(c('gapminder','ggplot2','gganimate','gifski'))

library(gapminder)
library(ggplot2)
library(gganimate)
library(gifski)

##Overall graph
ggplot(data=heart) +
  aes(totChol, age, size = 5, colour = ctr) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors)+
  scale_y_log10()+
  scale_size(range = c(2, 12))+
  labs(title = 'Total Cholestrol X Age', x = 'Total Cholestrol', y = 'Age')

#Total Cholestrol X Age by year
p1 <-ggplot(data=heart) +
  aes(totChol, age, size = 5, colour = ctr) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors)+
  scale_y_log10()+
  scale_size(range = c(2, 12))+
  facet_wrap(~ctr, ncol=6)+
##Annimate the plot
  labs(title = 'Year: {frame_time}', x = 'Total Cholestrol', y = 'Age')+
  transition_time(yr_cat) +
  ease_aes('linear')

ggsave('plot_Total_cholestrol_by_age.png', width=8, height=8)

animate(p1)
anim_save('Plot_TotalChol_Age_bycountry_by year')
```

Sample Data

male	age	diabetes	totChol	sysbp11	sysbp12	diaBP	BMI	heartRate	glucose	TenYearCHD	ctr	yr_cat
1	39	0	195	106	14	70	26.97	80	77	0	China	1990
0	46	0	250	121	15	81	28.73	95	76	0	Thailand	1990
1	48	0	245	127.5	-7.5	80	25.34	75	70	0	Japan	1990
0	61	0	225	150	-20	95	28.58	65	103	1	South Cor	1990
0	46	0	285	130	10	84	23.1	85	85	0	China	1991
0	43	0	228	180	-20	110	30.3	77	99	0	Thailand	1991
0	63	0	205	138	18	71	33.11	60	85	1	Japan	1991
0	45	0	313	100	-20	71	21.68	79	78	0	South Cor	1991
1	52	0	260	141.5	19	89	26.36	76	79	0	Australia	1991
1	43	0	225	162	-40	107	23.61	93	88	0	United Sta	1991
0	50	0	254	133	-13	76	22.91	75	76	0	Singapore	1991
0	43	0	247	131	-11	88	27.64	72	61	0	Malaysia	1991

Output



Was the conference successful?

What did the conference consist of?

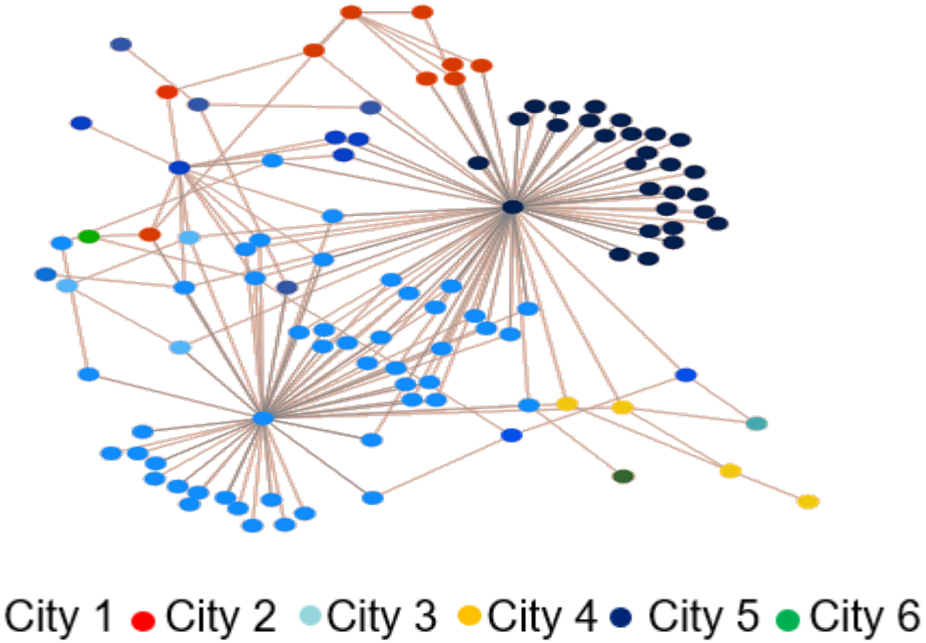
People connects and knowledge sharing sessions. Let us visualize the people connect.

Conference Execution

The conference was hosted by the internal data mining group.

Gauging Efficacy

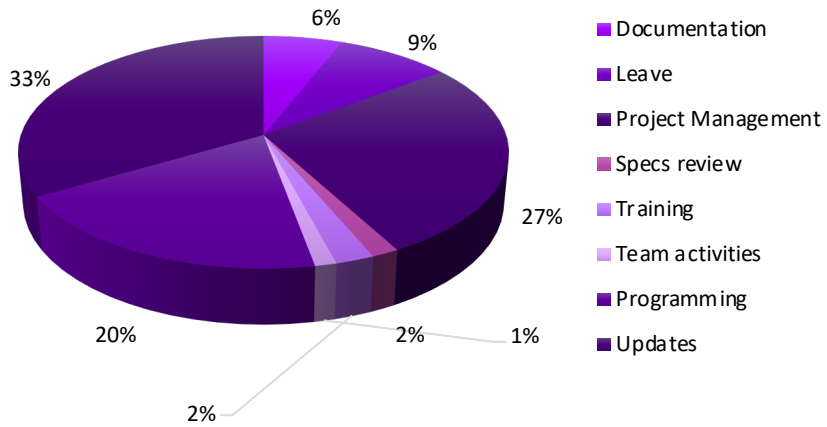
The conference's efficacy was gauged by asking the participants to provide the email addresses of the connections they made. This was then presented using a link diagram in Power BI.



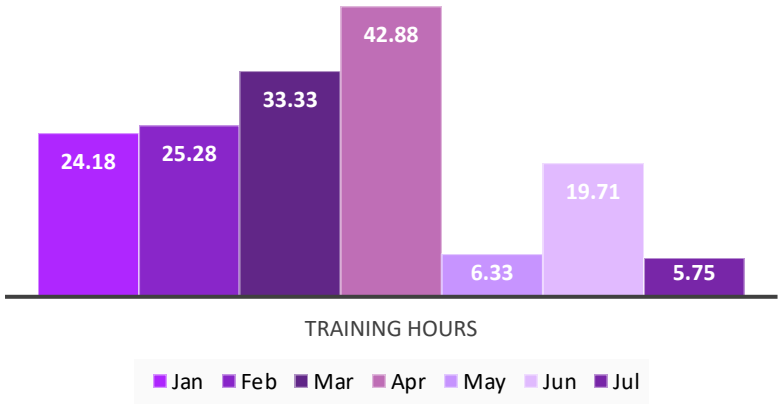
Here is a look at how these connections were forged -

Dashboard

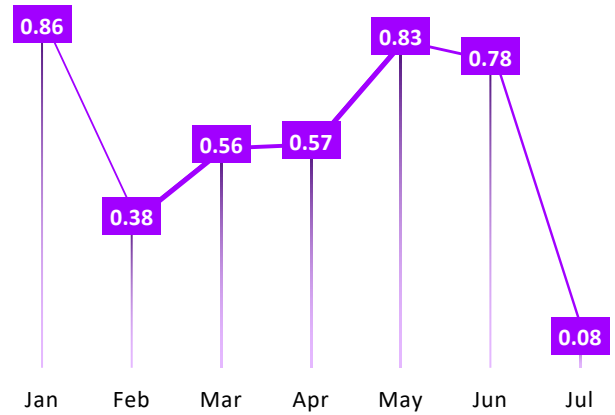
Activities



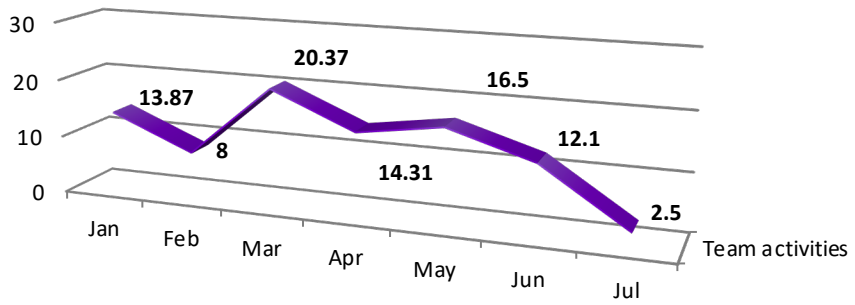
Training



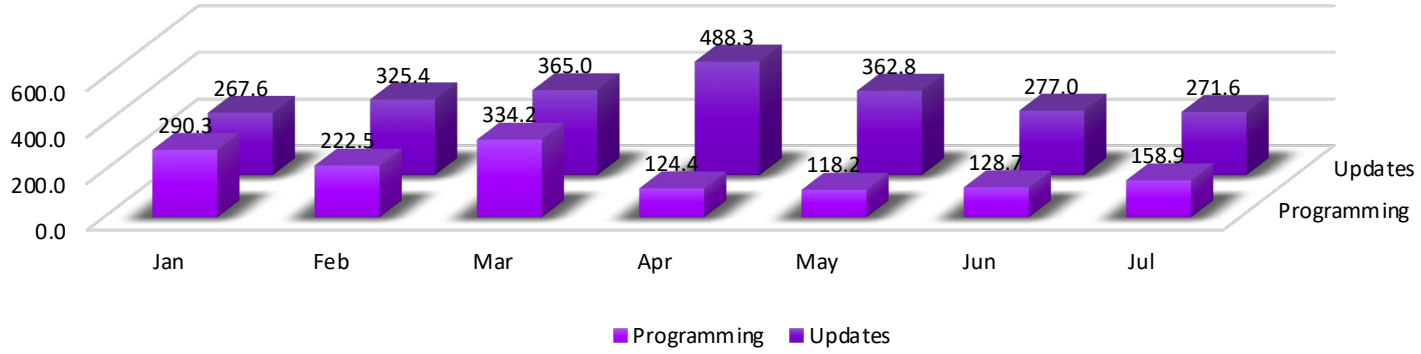
Leave



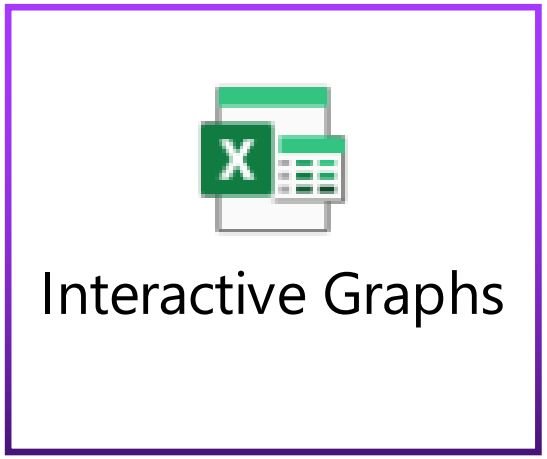
Fun activities



Programming and Updates



Interactive graph

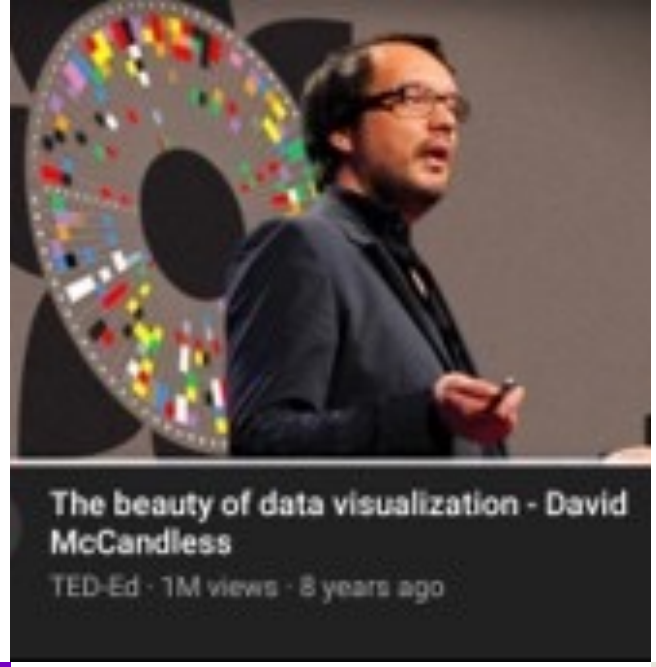


Check out the interactive graph in this workbook. Select different studies, and you will notice that the categories change accordingly.

Similar charts can be configured in Power BI

References

Click the links in the boxes to access the reference content.



01

The beauty of data visualization - David McCandless - YouTube

02

R: Documentation (r-project.org)

03

CRAN: Manuals (r-project.org)

04

Clinical Data Visualization: The Current State and Future Needs

Thank You