DATA 601: INTRODUCTION TO DATA SCIENCE PROJECT REPORT

A. SPANDANA

The reason I chose the question and dataset

TED (**Technology, Entertainment, Design**) is a media organization which posts talks online for free distribution, under the slogan "ideas worth spreading". TED was founded in February 1984 as a conference, which has been held annually since 1990. TED's early emphasis was technology and design, consistent with its Silicon Valley origins, but it has since broadened its focus to include talks on many scientific, cultural, and academic topics I've always been fascinated by TED Talks and the immense diversity of content that it provides for free. I was also thoroughly inspired by a TED Talk that visually explored TED Talks stats and I was motivated to do the same thing, albeit on a much less grander scale.

As a student, I occasionally watch TED Talks simply because they deliver thought-provoking ideas in an entertaining way. The recommendations given on the TED Talks website (www.ted.com/talks) are also useful, as they guide the audience toward other videos treating similar subjects, possibly offering contrasting perspectives. Many "institutions now offer open access to their intellectual property TED talks are a great way to improve your general knowledge. They will not make you knowledgeable in any specific field, but you will have contact with content from a huge diversity of subjects, directly from the insiders. They are a constructive form of entertainment. Presenters, who are passionate experts, speak with such energy and momentum. Their enthusiasm is contagious! They boost creativity. This is probably the main reason why I watch them. If you like your field or profession you tend to focus a lot on your tasks or the subject you're learning, researching or working on. Watching successful or inspired ideas, projects and solutions from the people that created or executed them, in very different fields, gives you different perspectives that you can apply within your own knowledge or experience, in your own projects They broaden your perspectives. By watching presentations on topics, you never imagined existed or had any attention, you start to compare you own problems and solutions to what exists out there and question yourself about what you really know. I usually enjoy TED talks because of the engaging style of the speakers. Each speaker has a different style which, is interesting, but they accomplish the same goal: captivate the audience.

2. Data Description:

These datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. The TED main dataset contains information about all talks including number of views, number of comments, descriptions, speakers and titles. The TED transcripts dataset contains the transcripts for all talks available on TED.com. The data has been scraped from the official TED Website and is available under the Creative Commons License.

Column name	Column Description
name	The official name of the TED Talk. Includes the title and the speaker.
num_speaker	The number of speakers in the talk

published_date	The Unix timestamp for the publication of the talk on TED.com
ratings	A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
related_talks	A list of dictionaries of recommended talks to watch next
speaker_occupation	The occupation of the main speaker
tags	The themes associated with the talk
title	The themes associated with the talk
url	The themes associated with the talk
views	The themes associated with the talk

Questions that can be answered:

Which are the most viewed and most favorited Talks of all time? Are they mostly the same? What does this tell us?
I find this interesting to know which the most favorite topic of all the time since 2011.This answer provides with the series of the favorite talks with the highest number of views and why it attracted the most viewers

What kind of topics attract the maximum discussion and debate (in the form of comments)?
The analysis of this question provides the information on what kind of discussion attracts the most topics and what does it target to. There are various topics on which ted talks takes place. It includes general knowledge, medicine, agriculture, education etc. I find it very fascinating how ted talks inspire people. Many famous personalities come forward and will share their experiences with the audience which is inspiring and motivation. I think it is a great medium to connect with people to inspire them. I think ted talks are great as they reach millions of people and influence every kind of a person from different places the answer provides the which of the talks from 2011 attracted most of the audience

Which months and years are most popular among TED and TEDx chapters?
This additional analysis provides with the number of talks that took place the most in a month. And on which days the most popular talks take place.

Which themes are most popular amongst TEDsters?
There are many themes under which the ted talks take. This answer benefits many people targeted to specific themes like business, innovation etc. This clearly provides the information on what themes attract the people under specific fields

Which speaker made the most appearances in tedtalks?
The answer to this question gives information on how many appearances have been made by the speakers, and who made the maximum appearances

Which profession made the most appearances?
Which profession has made the most appearances is answered by this question.

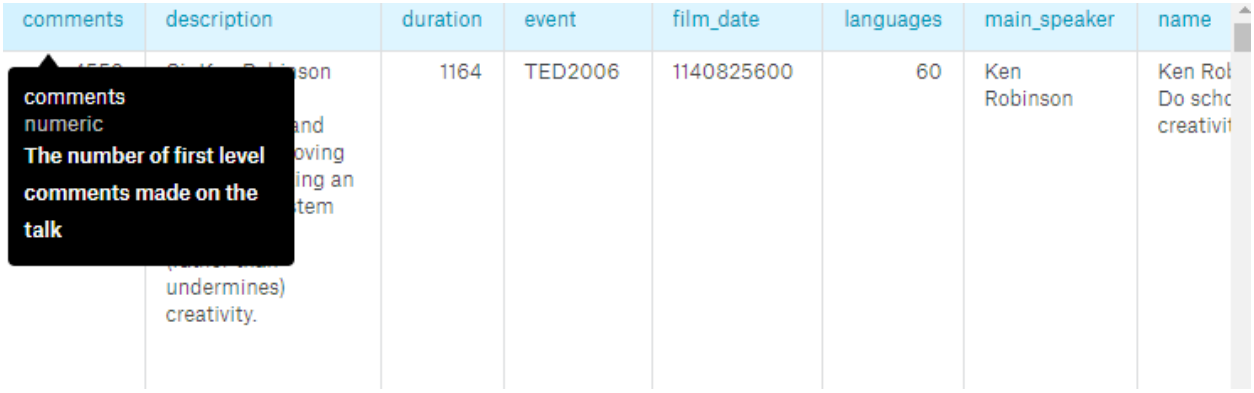
Analysis on Duration
Which Ted Talk is the longest and which ted talk is the shortest based on the seconds.

Analysis on Speakers
Which ted talk is having the most number of speakers

Analysis on Languages
Which ted talk was published in more languages

DATA CLEANING

- The data was provided from 2006 to September 2017.to get the recent data I decided to use the data from 2011.as it would give the most recent information for this purpose I used the subset command to filter out the necessary data set for the analysis.



a<-subset(tedtalks,published_date>"2010-12-31")

main_speaker	film_date	published_date	event	title
Arianna Huffington	2010-12-07	2011-01-03	TEDWomen 2010	How to succeed? Get more sleep
Lesley Hazleton	2010-10-10	2011-01-04	TEDxRainier	On reading the Koran
Charles Limb	2010-11-05	2011-01-05	TEDxMidAtlantic	Your brain on improv
Deborah Rhodes	2010-12-08	2011-01-06	TEDWomen 2010	A test that finds 3x more breast tumors, and wh
Neil Pasricha	2010-09-30	2011-01-07	TEDxToronto 2010	The 3 A's of awesome
Jody Williams	2010-12-08	2011-01-10	TEDWomen 2010	A realistic vision for world peace

- The published date is the published date of the talk. The subset command reads better from left and right and it is shorter, the data is filtered out from 2011.

The film_date and published_date is in the form of timestamp format. To get that into human readable format

	duration	event	film_date	lan
undly...	1164	TED2006	1140825600	
'An In...	977	TED2006	1140825600	
t tech...	1286	TED2006	1140739200	
activ...	1116	TED2006	1140912000	
e dra...	1190	TED2006	1140566400	

tedtalks1$film_date<-as.Date(as.POSIXct(val, origin="1970-01-01"))
tedtalks1$published_date<-as.Date(as.POSIXct(value, origin="1970-01-01"))

	film_date	l
	2006-02-25	
	2006-02-25	
	2006-02-24	
	2006-02-26	
	2006-02-22	

It is the direct method to return the data format.

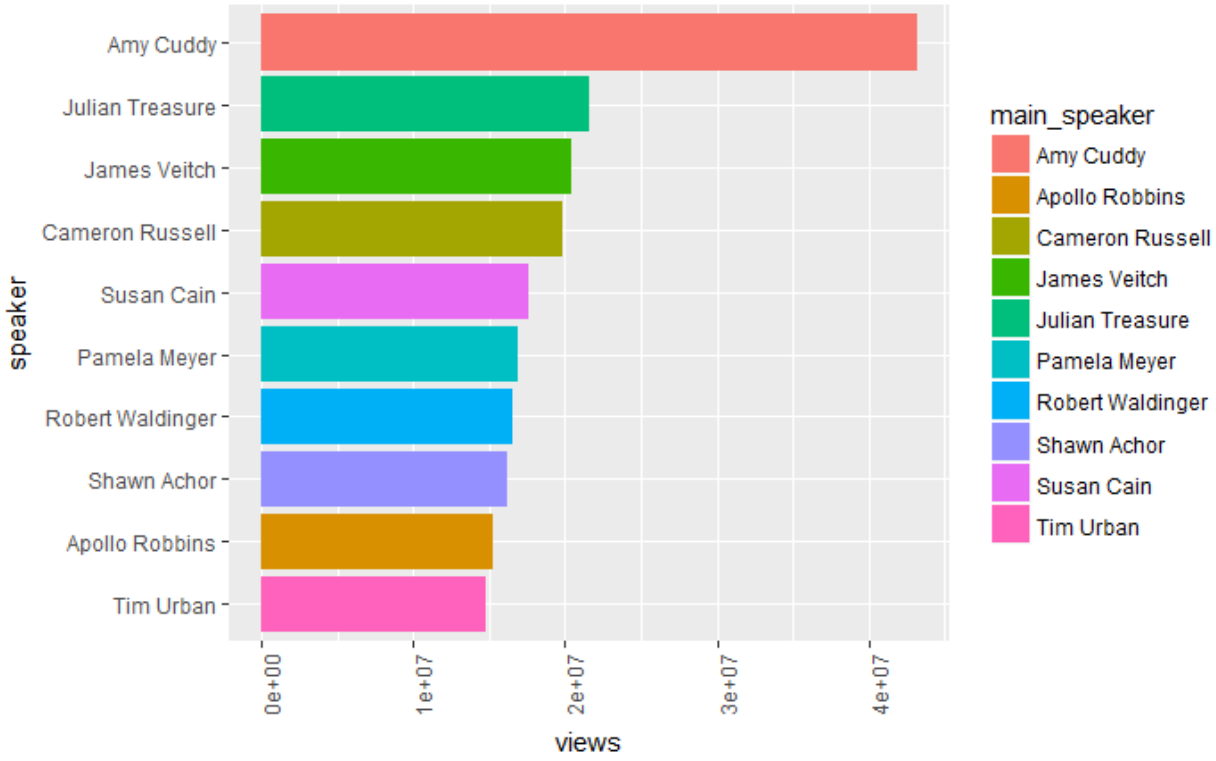
- I separated the date column into month day and year to perform analysis on the most repeated talks of all the time.

DATA ANALYSIS

1. MOST VIEWED TALK OF ALL THE TIME

```
tedtalks<-read.csv("C:\\Users\\spandanaadulla\\Documents\\data 601
project\\tedtalks\\ted_main.csv")
View(tedtalks)
value<-tedtalks$film_date
tedtalks$film_date<-as.Date(as.POSIXct(value,origin="1970-01-01"))
value<-tedtalks$published_date
tedtalks$published_date<-as.Date(as.POSIXct(value,origin="1970-01-01"))
tedtalks<-subset(tedtalks,published_date>"2010-12-31")
viewedtalks = tedtalks %>% arrange(desc(views)) %>% head(10)
View(viewedtalks)
ggplot(viewedtalks,aes(x=reorder(main_speaker,views), y=views,fill=main_speaker))+
geom_bar(stat="identity")+coord_flip()+labs(x="speaker")+theme(axis.text.x=element_t
```

ext(angle=90))



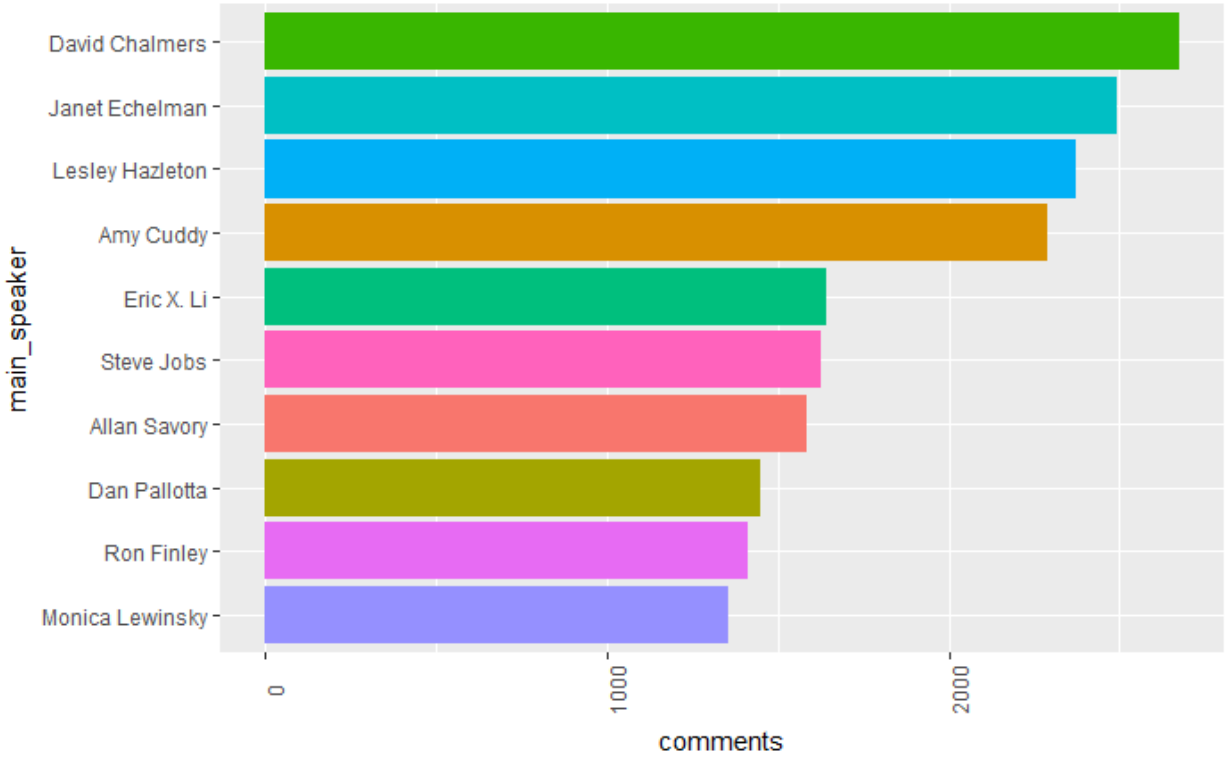
This plot is on the top ten most viewed talks from 2011 to September 2017.

Observations:

- Amy cuddy's: your body language may shape who you are is the most viewed with 43155405 published in October 2012

2. MOST COMMENTED OR MOST DISSCUSSED TALK OF ALL THE TIME

```
comments=tedtalks %>% arrange(desc(comments)) %>% head(10)
View(comments)
tedtalks1<-tedtalks%>%arrange(desc(comments))%>%head(10)
ggplot(tedtalks1, aes(x=reorder(main_speaker,comments), y=comments,
fill=main_speaker)) +geom_bar(stat = 'identity') +guides(fill=FALSE) +
labs(x="main_speaker")+theme(axis.text.x=element_text( angle=90))+coord_flip()
View(head(tedtalks1[c('title','main_speaker','event','name')]))
```



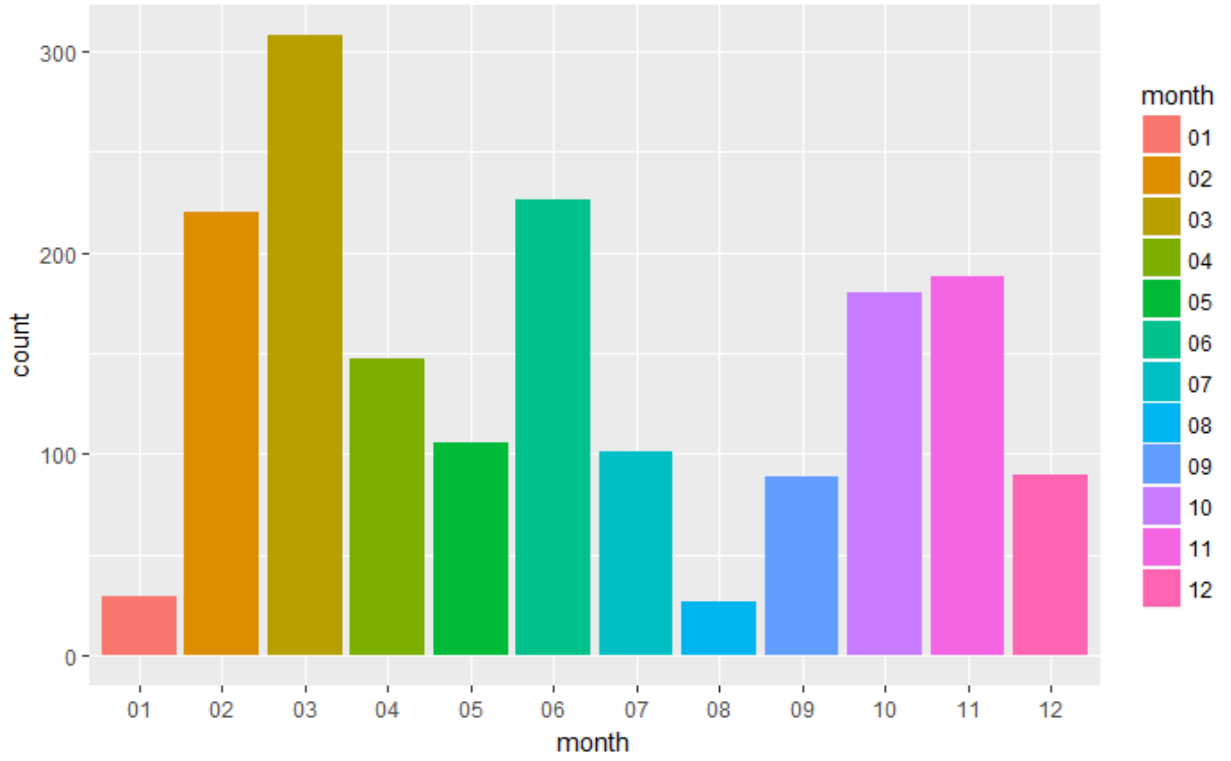
	title	main_speaker	event	name
1	How do you explain consciousness?	David Chalmers	TED2014	David Chalmers: How do you explain consciousness?
2	Taking imagination seriously	Janet Echelman	TED2011	Janet Echelman: Taking imagination seriously
3	On reading the Koran	Lesley Hazleton	TEDxRainier	Lesley Hazleton: On reading the Koran
4	Your body language may shape who you are	Amy Cuddy	TEDGlobal 2012	Amy Cuddy: Your body language may shape who you are
5	A tale of two political systems	Eric X. Li	TEDGlobal 2013	Eric X. Li: A tale of two political systems
6	How to live before you die	Steve Jobs	Stanford University	Steve Jobs: How to live before you die

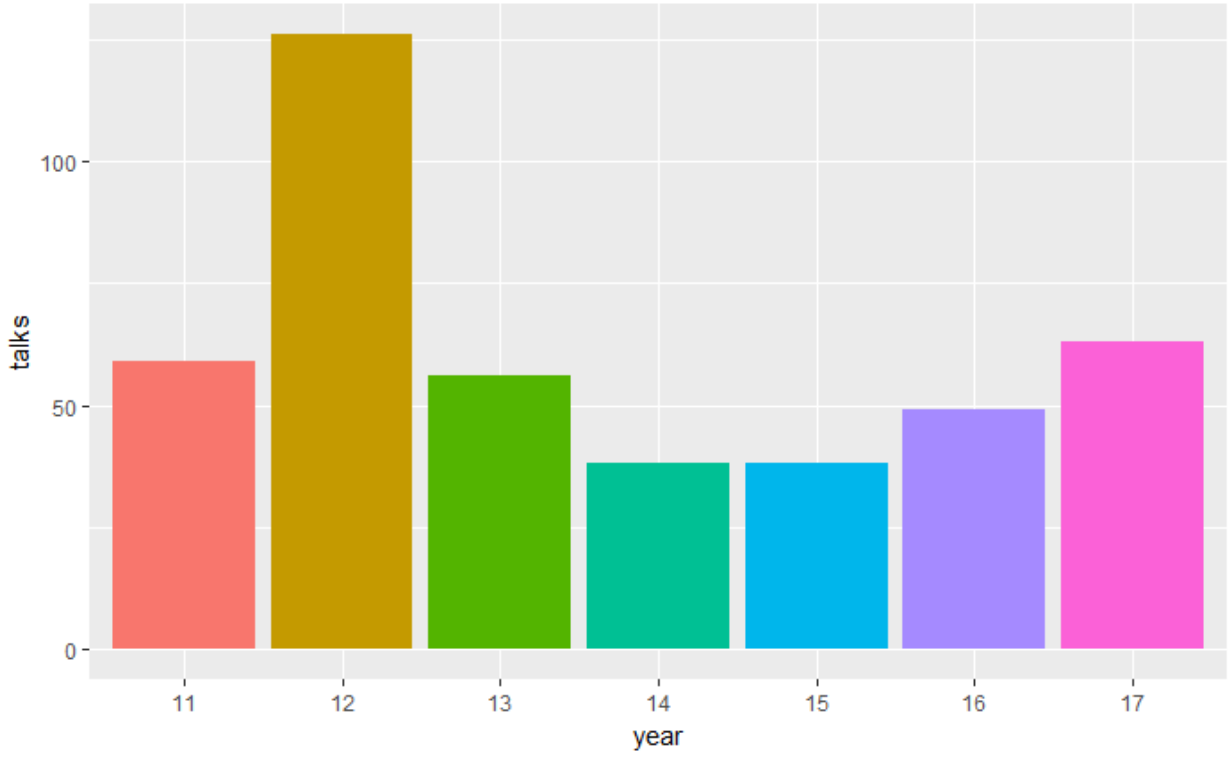
Observations:

- David Chalmers "how do you explain consciousness?" which was held in TED2014 is having 2673 comments.

3.TEDTALKS COUNT BY YEAR AND MONTH

```
tedtalks$month<-format(as.Date(tedtalks$film_date,format="%d/%m/%Y"),"%m")
tedtalks$date<-format(as.Date(tedtalks$film_date,format="%d/%m/%Y"),"%d")
tedtalks$year<-format(as.Date(tedtalks$film_date,format="%d/%m/%Y"),"%y")
tedtalks$monthp<-
format(as.Date(tedtalks$published_date,format="%d/%m/%Y"),"%m")
tedtalks$date<-format(as.Date(tedtalks$published_date,format="%d/%m/%Y"),"%d")
tedtalks$year<-format(as.Date(tedtalks$published_date,format="%d/%m/%Y"),"%y")
ggplot(tedtalks, aes(x=month,fill=month)) + geom_bar()
tedtalks2 <- tedtalks[grep("TEDx", tedtalks$event),]
yearlytalks<- data.frame(table(tedtalks2$year))
colnames(yearlytalks) <- c("year", "talks")
ggplot(yearlytalks, aes(x=year, y=talks,fill=year)) + geom_bar(stat='identity') +
guides(fill=FALSE)
```





Observations:

- It can be observed that 2012 is the year and march is the month with highest number of talks comparatively.

4.MOST APPEARENCES BY SPEAKERS

```
speaker <- data.frame(table(tedtalks$main_speaker))
colnames(speaker) <- c("main_speaker", "appearances")
speaker<- speaker %>% arrange(desc(appearances))
head(speaker)
```

	main_speaker <fctr>	appearances <int>
1	Marco Tempest	6
2	Juan Enriquez	4
3	Andrew Solomon	3
4	Bill Gates	3
5	Christopher Soghoian	3
6	Dan Ariely	3

The table shows the speakers who have made the most number of appearances from 2011.

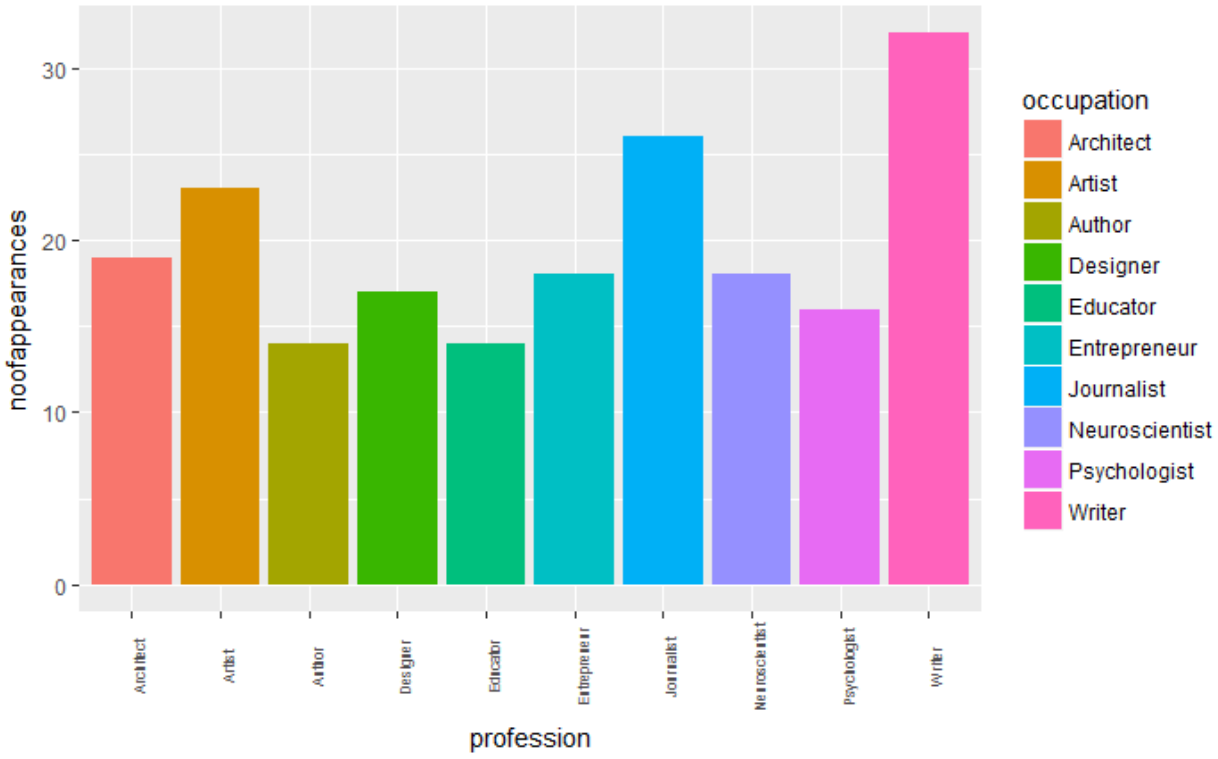
Observations:

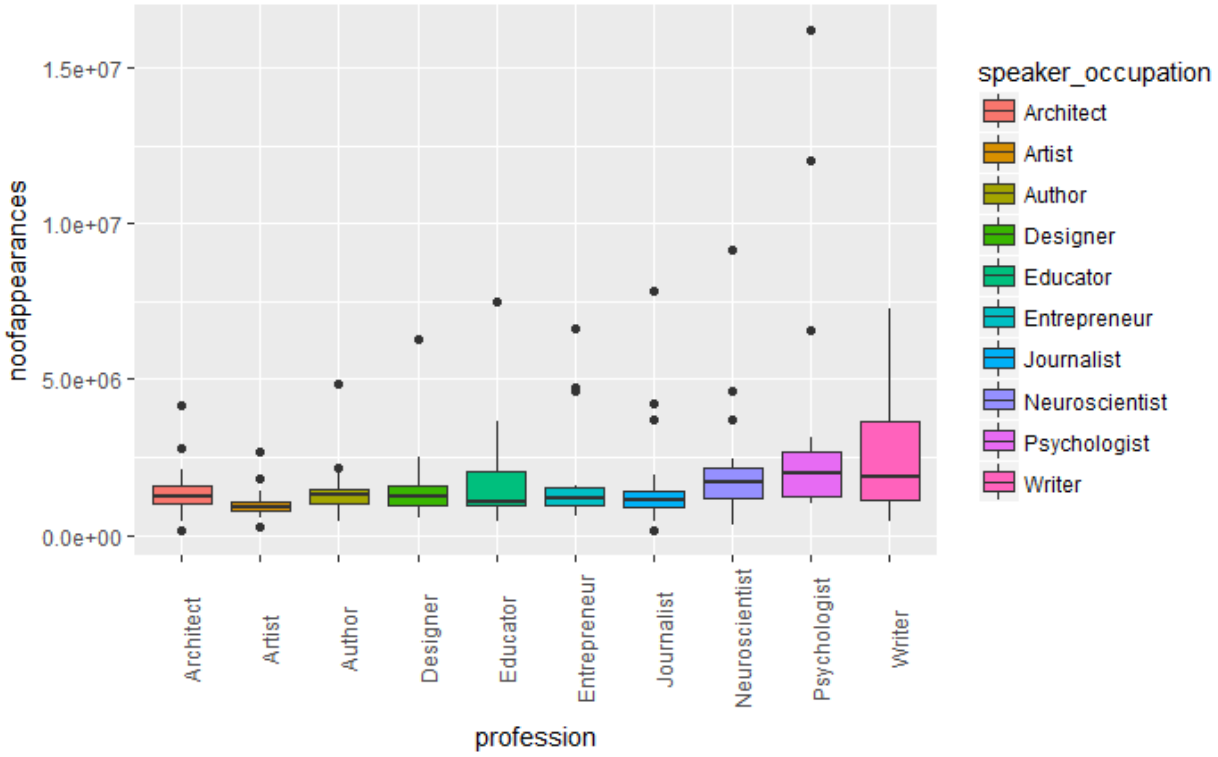
- Marco Tempest who is a swiss magician from New York has made six appearances

5.APPEARENCES ON OCCUPATION

```
occupation_tedtalks <- data.frame(table(tedtalks$speaker_occupation))
colnames(occupation_tedtalks) <- c ("occupation", "appearancesofthatoccupation")
occupation_tedtalks <- occupation_tedtalks %>%
arrange(desc(appearancesofthatoccupation))
head(occupation_tedtalks, 10)
ggplot(head(occupation_tedtalks,10), aes (x=occupation, y=appearancesofthatoccupation,
fill=occupation)) + geom_bar(stat=" identity") +theme (axis.text.x=element_text( si
ze=6,angle=90))+labs(x="profession",y="noofappearances")
tedtalks_common_occ <- tedtalks[tedtalks$speaker_occupation %in%
head(occupation_tedtalks$occupation, 10), ]
ggplot (tedtalks_common_occ, aes (x=speaker_occupation, y=views,
fill=speaker_occupation)) +
geom_boxplot()+theme(axis.text.x=element_text(angle=90))+labs(x="profession",
y="noofappearances")
```

	occupation <fctr>	appearancesofthatoccupation <int>
1	Writer	32
2	Journalist	26
3	Artist	23
4	Architect	19
5	Entrepreneur	18
6	Neuroscientist	18
7	Designer	17
8	Psychologist	16
9	Author	14
10	Educator	14





Observations:

- Writer is the profession from where most of the speakers made appearances.
- Artists the profession which has made the least appearances over the time.

6.YEAR WHICH HAD MOST OF THE TED TALKS

```
event_tedtalks <- data.frame(table(tedtalks$event))
colnames(event_tedtalks) <- c("event name", "talks")
event_tedtalks <- event_tedtalks %>% arrange(desc(talks))
head(event_tedtalks, 10)
```

	event name <fctr>	talks <int>
1	TED2014	84
2	TED2013	77
3	TED2016	77
4	TED2015	75
5	TED2011	70
6	TEDGlobal 2012	70
7	TEDGlobal 2011	68
8	TED2017	67
9	TEDGlobal 2013	66
10	TED2012	65

The basic difference between Ted Global events takes place at a global platform and ted events take place nationally

Observations:

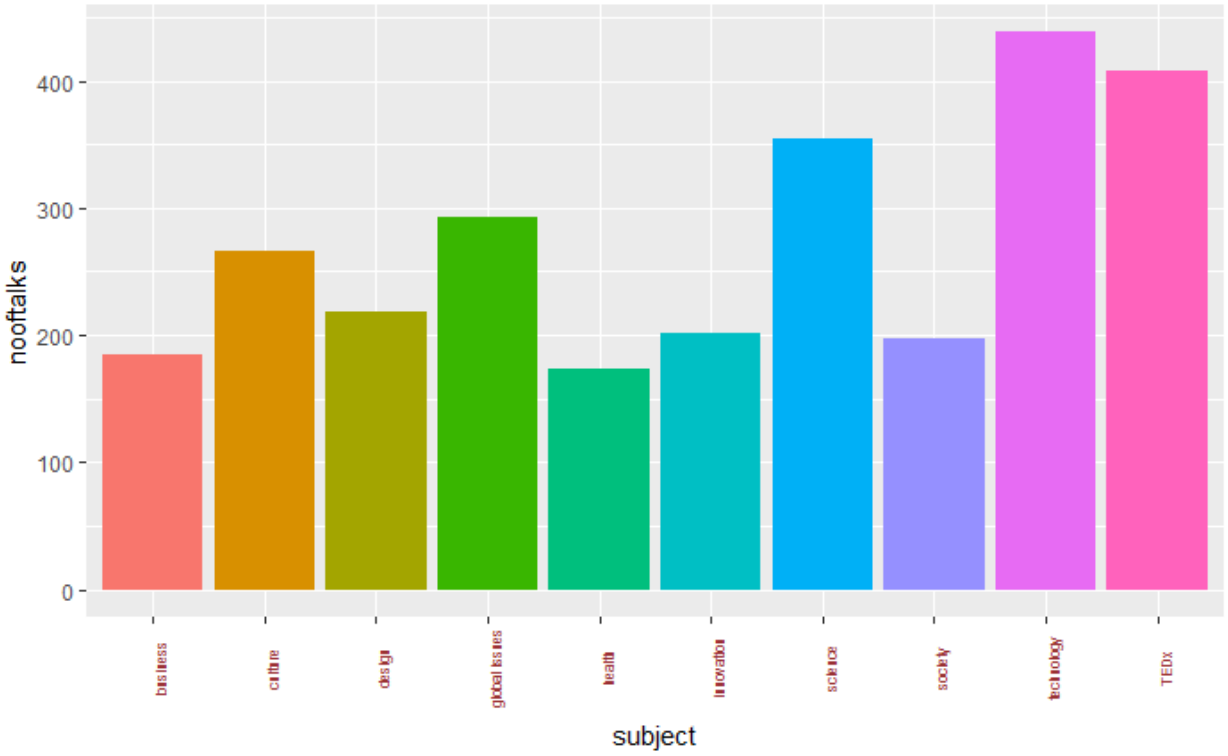
- The data frame shows that Ted talks in 2012 had the most number of talks

7.THEME TED TALKS

```
rep.row <- function(r, n)
{
colwise(function(x) rep(x, n))(r)
}
get_list <- function(s)
{
s <- as.character(s)
s <- strsplit(s, "\\[")[[1]][2]
s <- strsplit(s, "\\]")[[1]][1]
s <- strsplit(s, ",")[[1]]
s_list <- sapply(strsplit(s, """), function(x) x[2])
return(s_list)
}
result <- function(ix, tednooftalks)
{
tags <- get_list(tednooftalks$tags[ix])
latestteds <- rep.row(tednooftalks[ix, ], length(tags))
latestteds$subject <- tags
return(latestteds)
}
merge_bits <- function(a, b)
```

```
{
return(rbind(data.frame(a), data.frame(b)))
}
resulted <- lapply(1:nrow(tedtalks), function(x) result(x, tedtalks))
subject_tednooftalks <- Reduce(merge_bits, resulted)
rm(resulted)
subject<-head(subject_tednooftalks)
subjects = data.frame(table(subject_tednooftalks$subject))
colnames(subjects) <- c("subject", "nooftalks")
subjects <- subjects %>% arrange(desc(nooftalks))
subjects %>% head(10)
ggplot(head(subjects,10), aes(x=subject, y=nooftalks, fill=subject)) +
geom_bar(stat="identity") +
guides(fill=FALSE)+theme(axis.text.x=element_text(color="#993333", size=6,
angle=90))
```

	subject <fctr>	nooftalks <int>
1	technology	438
2	TEDx	408
3	science	354
4	global issues	293
5	culture	266
6	design	218
7	innovation	201
8	society	198
9	business	185
10	health	173

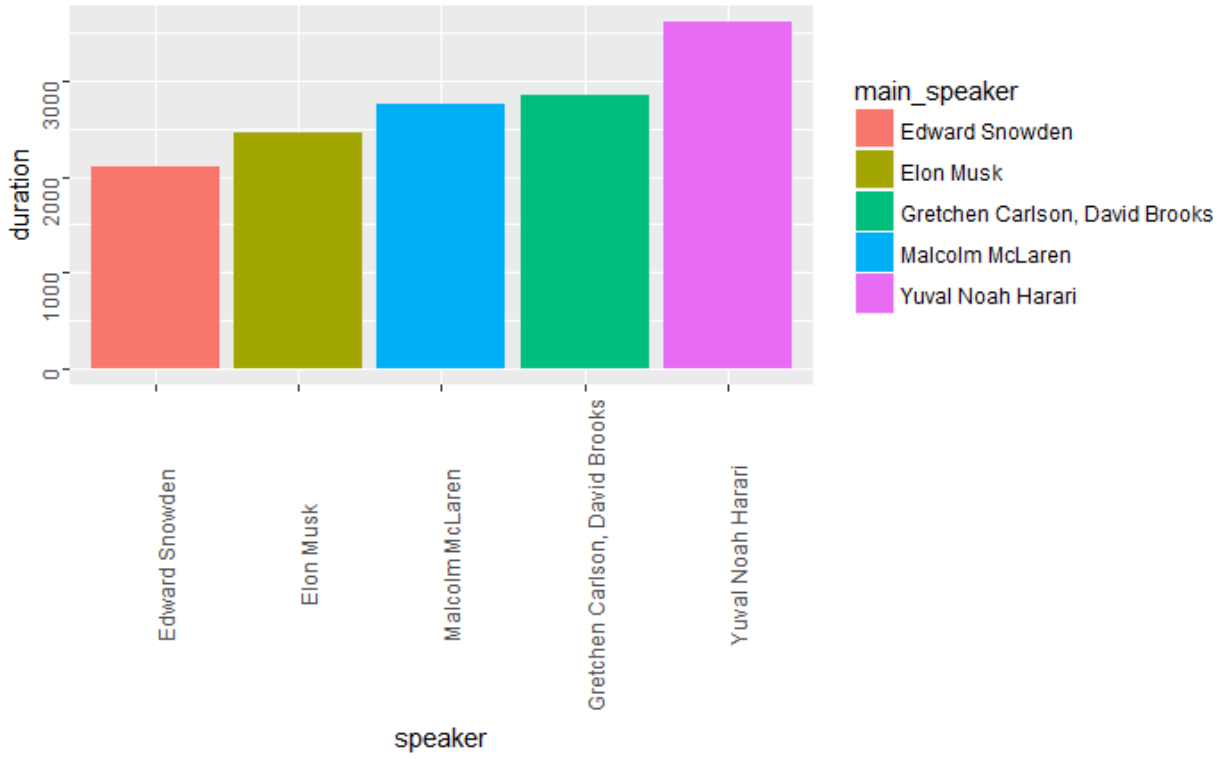


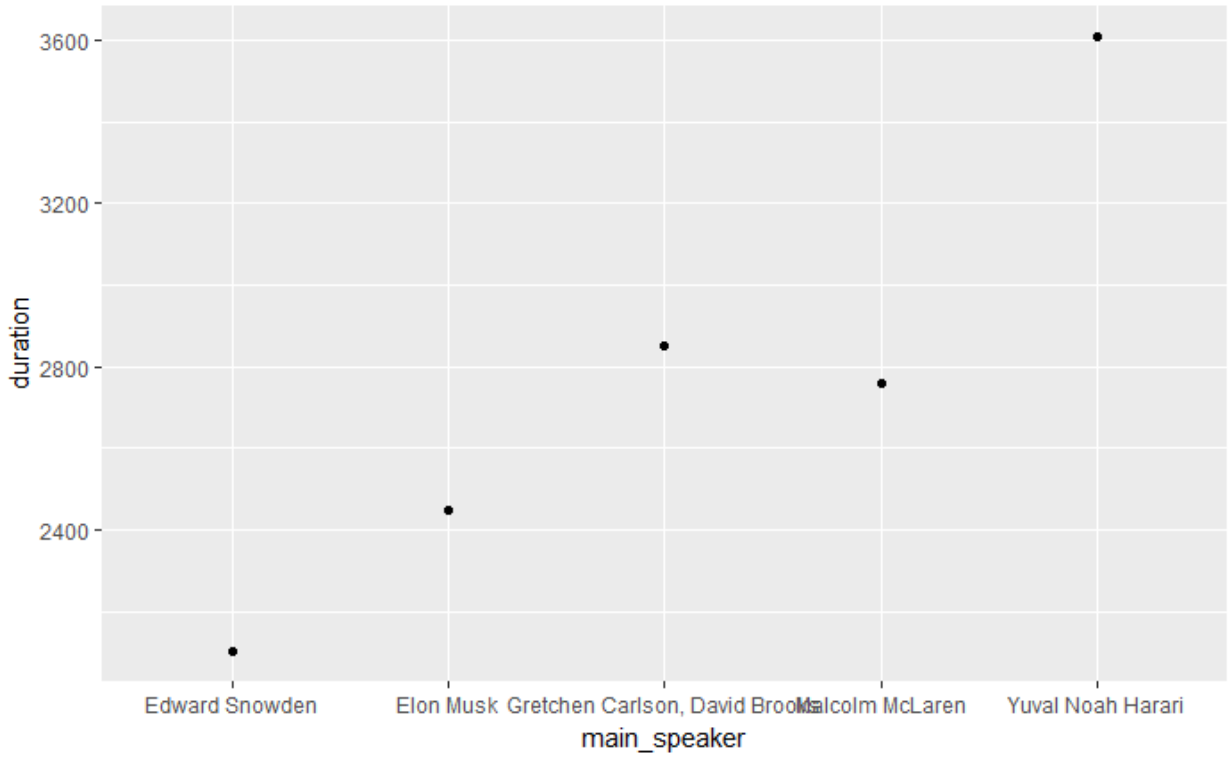
Observations:

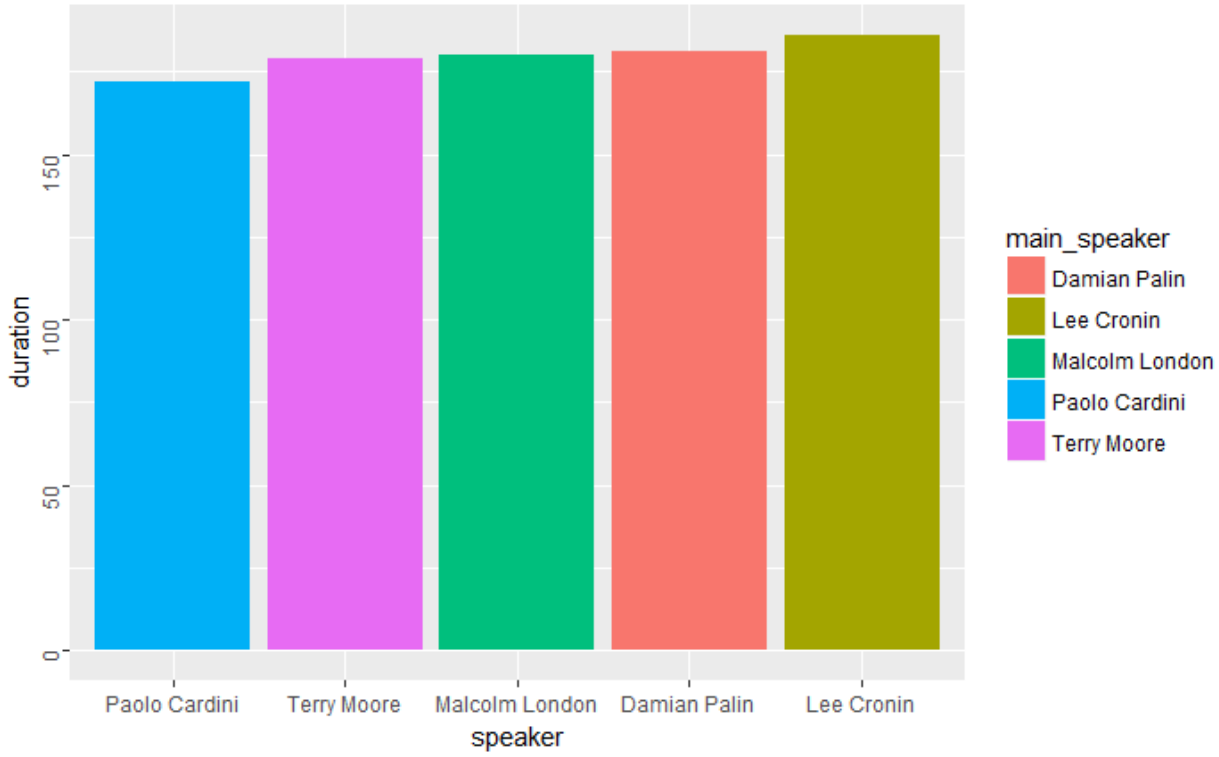
- Talks related to technology had more than 400 talks from all the time
- Talks related to health had the least number of ted talks which are 173

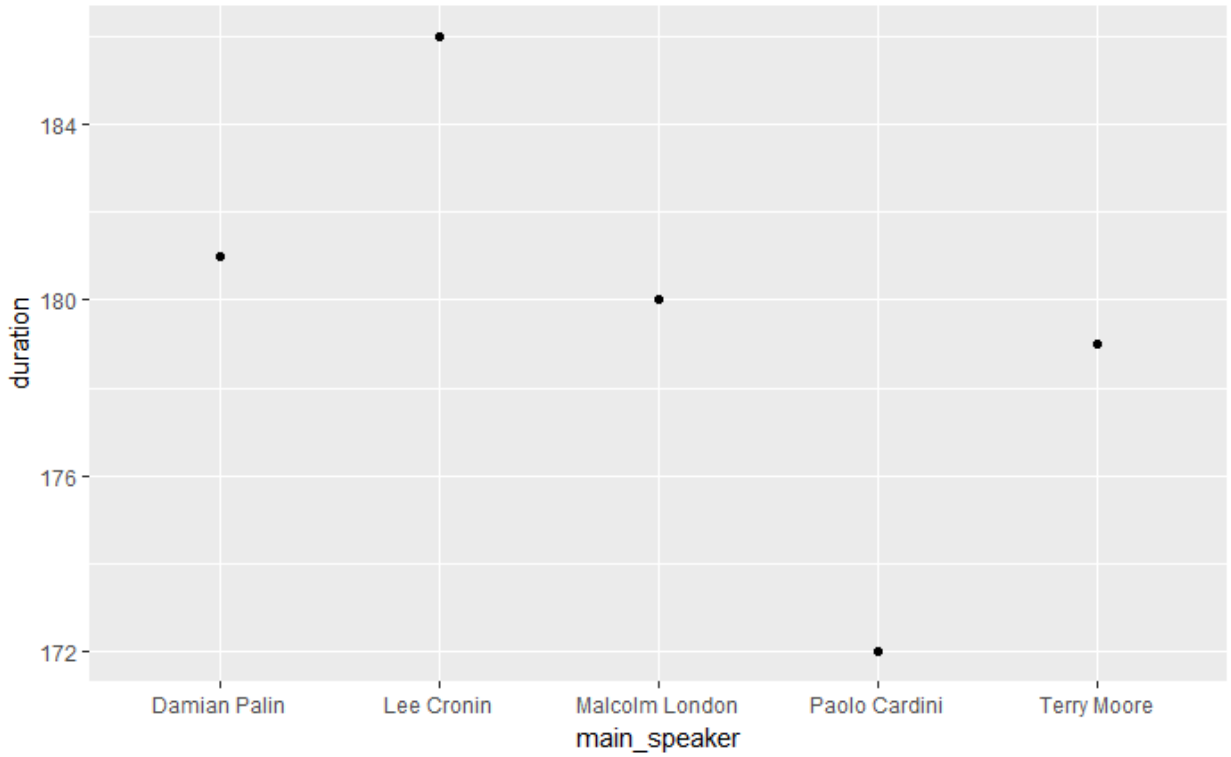
8.ANALYSIS ON DURATION

```
length<-tedtalks %>% arrange(desc(duration)) %>% head(5)
ggplot(length,aes(x=reorder(main_speaker,duration),y=duration,fill=main_speaker))+
geom_bar(stat="identity")+labs(x="speaker")+
theme(axis.text.y=element_text(angle=90),axis.text.x=element_text(angle=90))
ggplot()+layer(data=length,mapping =
aes(x=main_speaker,y=duration),geom="point",stat="identity",position="identity")
duration<-tedtalks$duration
length2<-tedtalks[order(duration),]%>%head(5)
View(length2)
ggplot(length2,aes(x=reorder(main_speaker,duration), y=duration,fill=main_speaker))+
geom_bar(stat="identity")+labs(x="speaker")+theme(axis.text.y=element_text(angle=90)
)
ggplot()+layer(data=length2,mapping =
aes(x=main_speaker,y=duration),geom="point",stat="identity",position="identity")
```




3600
3200
2800
2400
duration
Edward Snowden
Elon Musk
Gretchen Carlson, David Brooks
Malcolm McLaren
Yuval Noah Harari
main_speaker


150
100
50
0
duration
Paolo Cardini
Terry Moore
Malcolm London
Damian Palin
Lee Cronin
speaker
main_speaker
Damian Palin
Lee Cronin
Malcolm London
Paolo Cardini
Terry Moore



Observations:

- yuval noah Harari's is the longest talk: nationalism vs globalism with 3608 published on February 2015
- Paolo Cardini is the shortest talk Forget multitasking, try monotasking with 172 secs published in June 2012

9.ANALYSIS ON NUMBER OF SPEAKERS

```
numberofspeakers = tedtalks %>% arrange(desc(num_speaker)) %>% head(2)
numberofspeakers[,c("title","num_speaker","main_speaker")]
>
```

	title <fctr>	num_speaker <int>
1	A dance to honor Mother Earth	5
2	The interspecies internet? An idea in progress	4

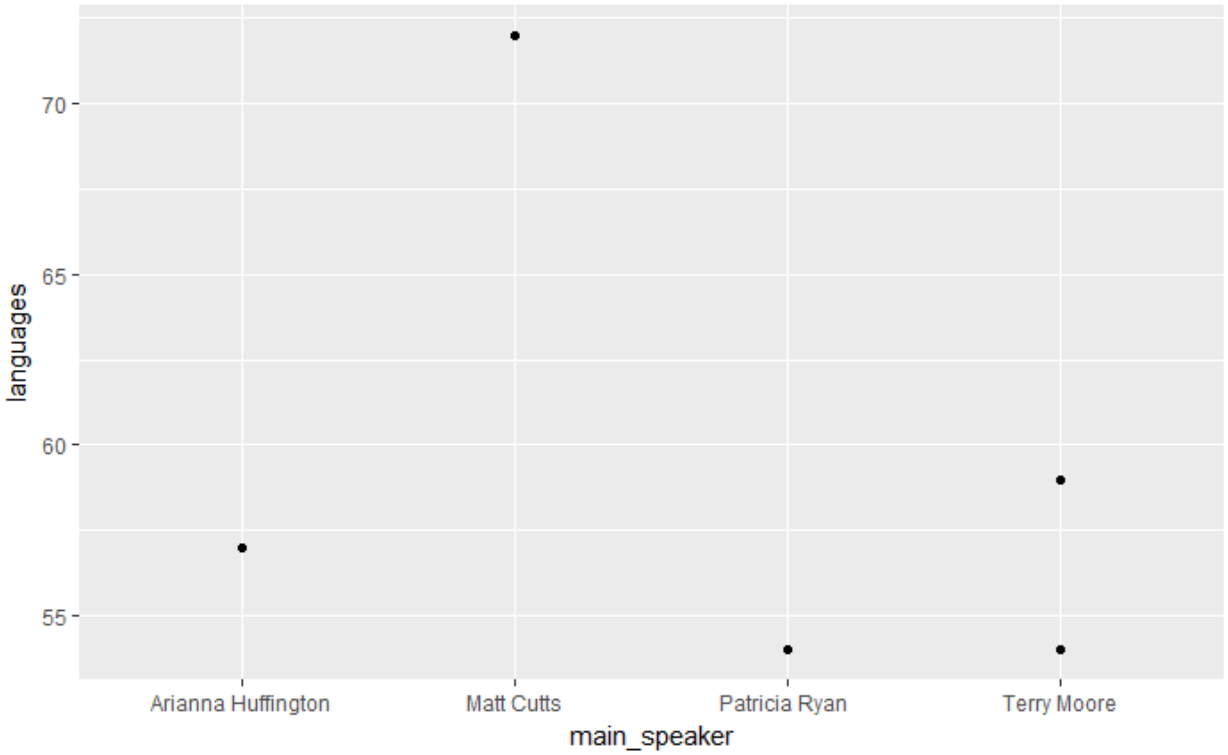
Observations:

- A dance to honor Mother Earth with five speakers followed by the interspecies internet? An idea in progress are the ted talks which are having more number of speakers

10.ANALYSIS ON LANGUAGES

```
numberoflanguages = tedtalks %>% arrange(desc(languages)) %>% head(5)
View(numberoflanguages)
```

```
ggplot()+layer(data=numberoflanguages,mapping =
aes(x=main_speaker,y=languages),geom="point",stat="identity",position="identity")
```



Observations:

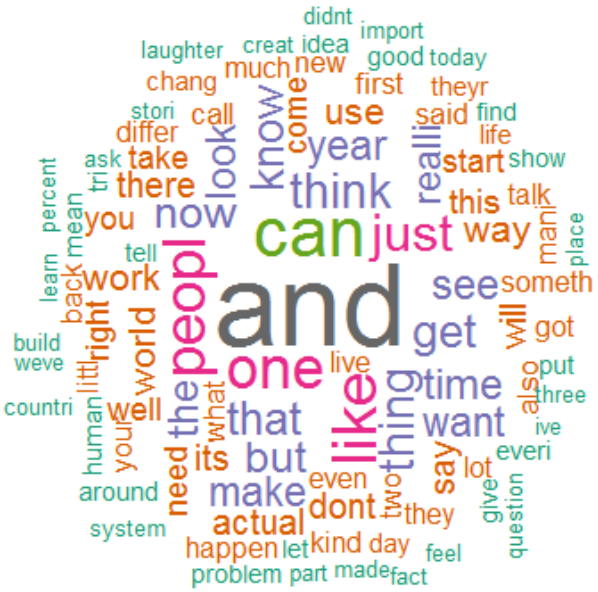
- Matt Cutts: Try something new for 30 days was published in 72 languages.

11.WORD CLOUD

```
transcripts<-read.csv("C:\\Users\\spandanaadulla\\Documents\\data 601
project\\tedtalks\\transcripts.csv")
texts <- transcripts$transcript
corpus <- Corpus(VectorSource(texts))
corpus <- tm_map(corpus, PlainTextDocument)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeWords, stopwords('english'))
corpus <- tm_map(corpus, stemDocument)
corpus <- Corpus(VectorSource(corpus))
dtm <- TermDocumentMatrix(corpus)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
max.words=100, random.order=FALSE, rot.per=0.35,
```

colors=brewer.pal(8, "Dark2"))

	word <fctr>	freq <dbl>
and	and	42887
can	can	24129
one	one	20275
like	like	19814
peopl	peopl	19527
just	just	16098
thing	thing	14545
think	think	14370
that	that	13963
get	get	13840



SUMMARY/CHALLENGES

I found it interesting and fascinated to work on ted talks I learnt about the word cloud and text mining by implementing the corpus function by implementing tm and word cloud packages.

CITATION:

https://www.kaggle.com/rounakbanik/ted-talks