Final Project Report: Predicting TED Talk Views

Kathleen Cachel

Load needed packages.

```
library(ggplot2) # Data visualisation
library(reshape2)
library(corrplot)
library(dplyr)
library(ggthemes)
library(anytime)
library(lubridate)
library(ggpubr)
library(tm)
library(tidyverse)
TedRaw <- read.csv(file="ted_main.csv", header=TRUE, sep=",")
```

Exploratory Data Analysis

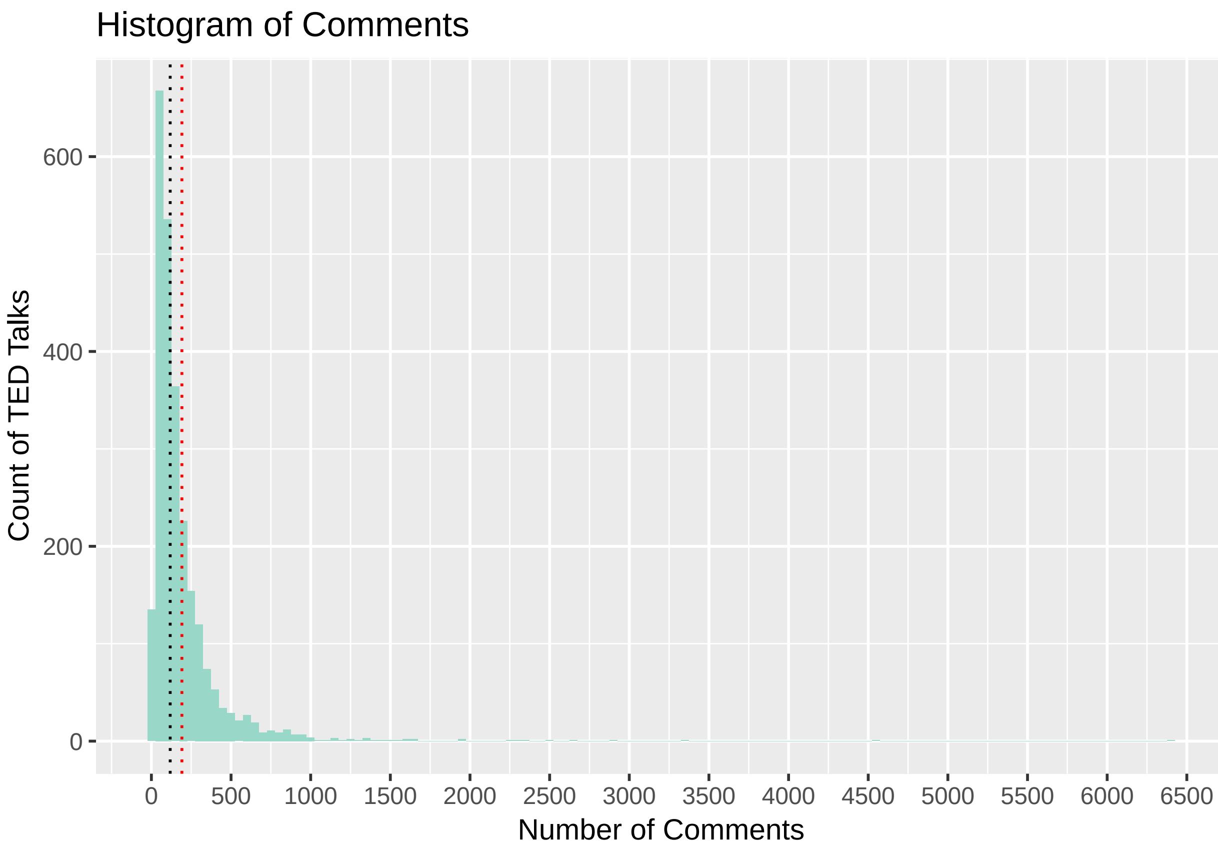
The TED Talk dataset contains entries for 2550 talks on the TED website and contains 17 variables (of which the predictor **views** is one). Some variables are useful as is and some require data munging or transformation. Below we work through the variables and outline their potential transformations.

Variable: comments

The variable **comments** is an interger denoting how many top level comments are posted to a talks site. Without knowing all that much it seems a reasonable hypothesis would be comments is a decent predictor of **views**.

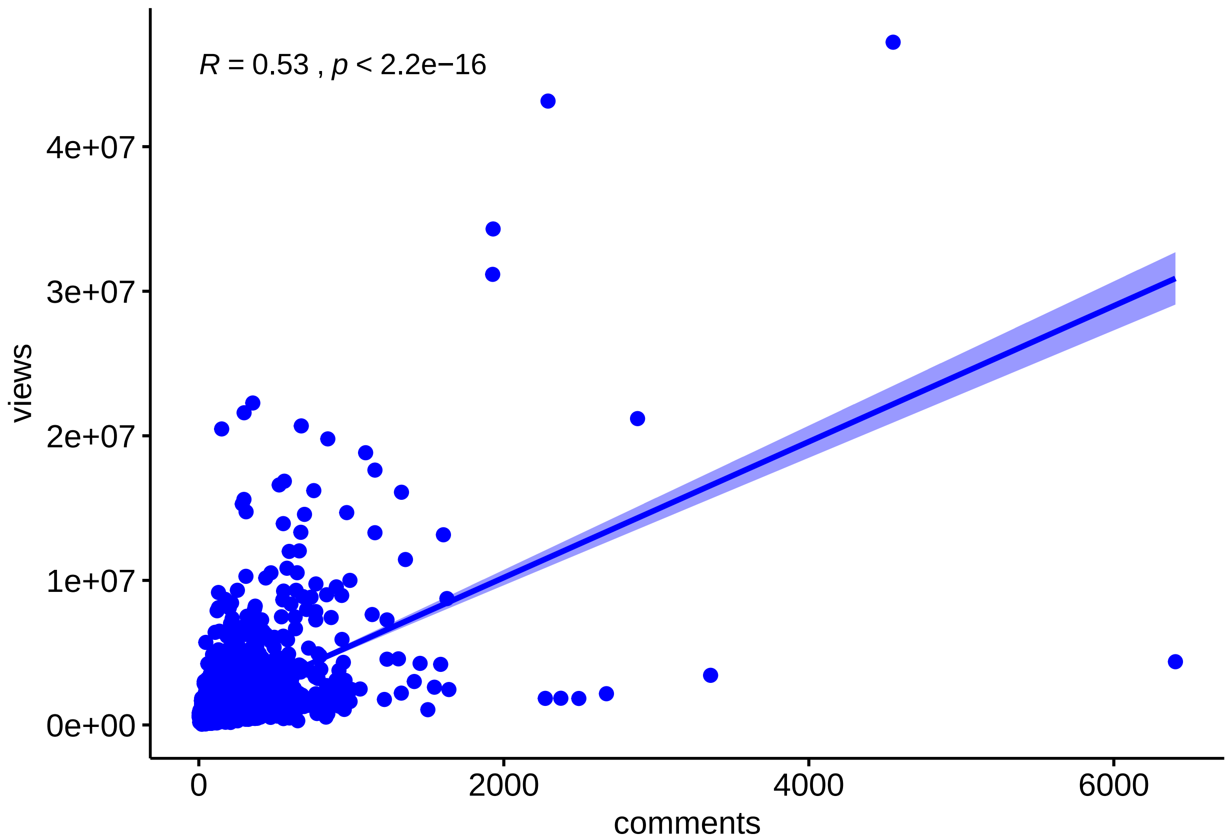
Below is a histogram graph binning TED talks by the number of comments on the talk's site. We can see the median (black) is left of the average (red), indicating it is right skewed.

```
#Number of Comments
#Histogram Graph
ggplot(data=TedRaw, aes(x=comments)) +
  geom_histogram(binwidth=50, fill = '#99d8c9')+
  geom_vline(aes(xintercept = mean(comments)),col='red',
             size=.5, linetype="dotted")+
  geom_vline(aes(xintercept = median(comments)),col='black',
             size=.5, linetype="dotted")+
  labs(x = "Number of Comments", y = "Count of TED Talks",
       title = "Histogram of Comments")+
  scale_x_continuous(breaks = pretty(TedRaw$comments, n = 22))
```



Below is a plot of the observation with **comments** on the X-axis and **views** on the Y-axis. We can see that the R value is 0.53 and the p value is significant. Based on this information, we will likely keep the **comments** variable in future models.

```
ggscatter(TedRaw, x = "comments", y = "views",
          add = "reg.line", conf.int = TRUE, color = "blue",
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "comments", ylab = "views")
```

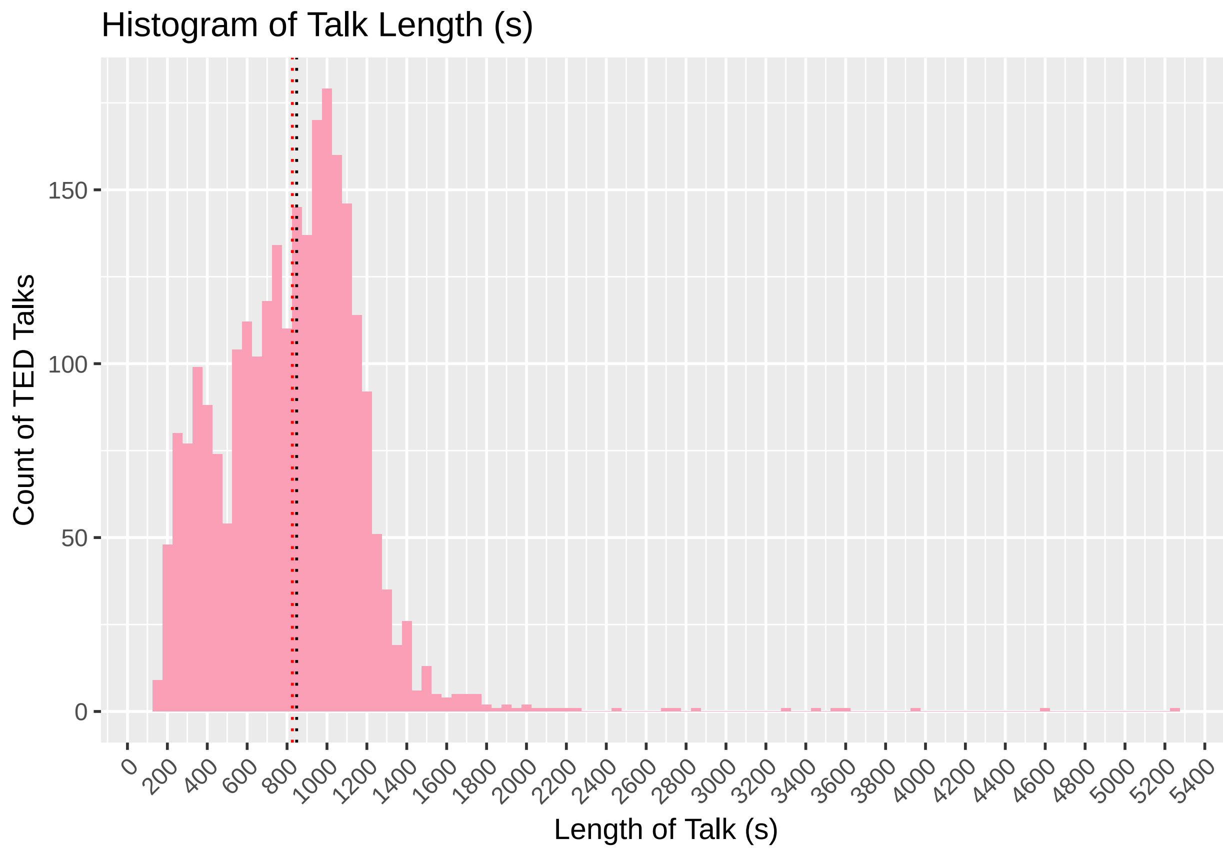


Variable: duration

The variable **duration** is an interger denoting how long a TED Talk is in seconds. Without knowing much we might assume that longer talks have less views since people may be less inclined to watch longer videos.

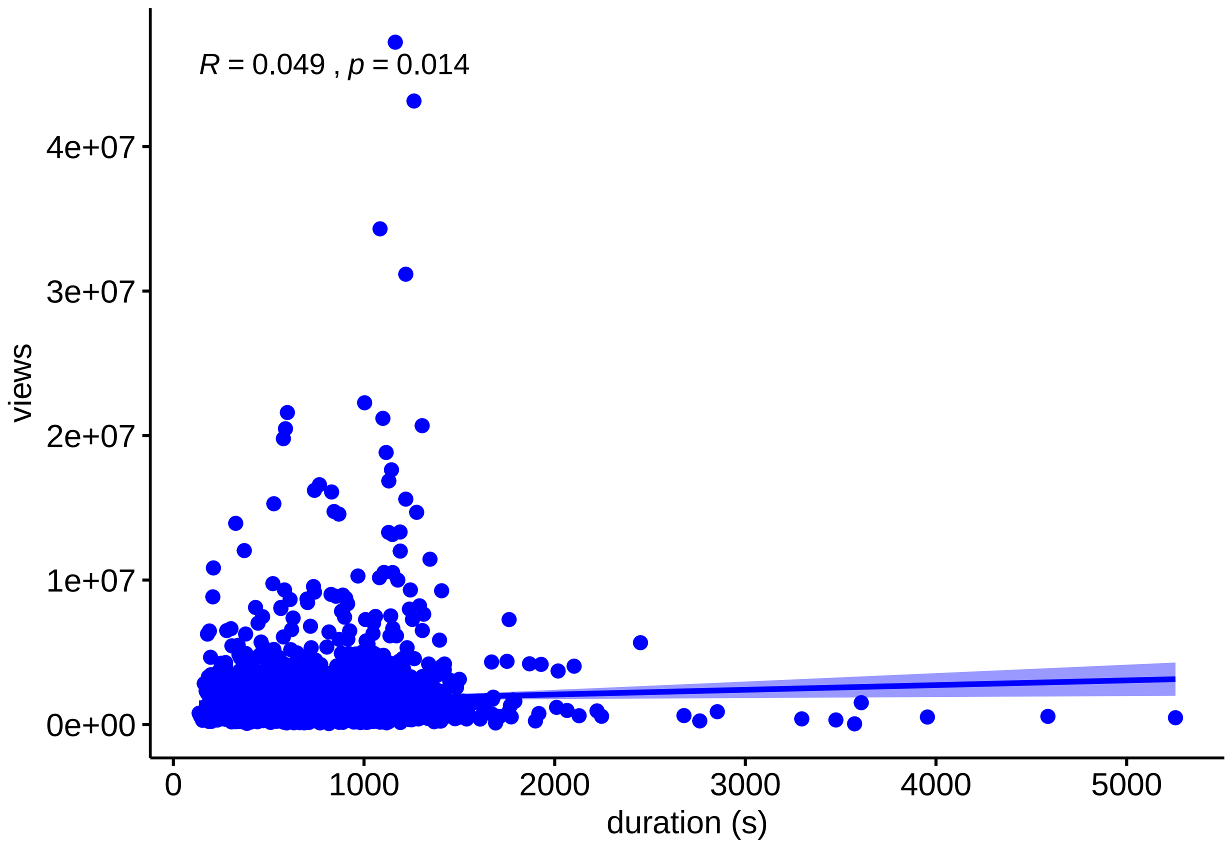
Below is a histogram graph binning TED talks by their length. We can see the median (black) is pretty close to the average (red). It looks like the majority of talks are under 1200 seconds, and the most are around 1000-1200.

```
#Duration of Talk
duration <- TedRaw %>%
  group_by(duration) %>%
  tally(name ="numTalks")

#histogram
ggplot(data=TedRaw, aes(x=duration)) +
  geom_histogram(binwidth=50, fill = '#fa9fb5')+
  geom_vline(aes(xintercept = mean(duration)),col='red',
             size=.5, linetype="dotted")+
  geom_vline(aes(xintercept = median(duration)),col='black',
             size=.5, linetype="dotted")+
  labs(x = "Length of Talk (s)", y = "Count of TED Talks",
       title = "Histogram of Talk Length (s)")+
  scale_x_continuous(breaks = pretty(TedRaw$duration, n = 22))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Is **duration** correlated with **views**? Not really.Below is a plot of the observation with **duration** on the X-axis and **views** on the Y-axis. We can see that the R value is pretty small (0.049) and the p value is significant. Based on this information, we will likely keep the **duration** variable in future models, but be mindful that it may have little effect.

```
ggscatter(TedRaw, x = "duration", y = "views",
          add = "reg.line", conf.int = TRUE, color = "blue",
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "duration (s)", ylab = "views")
```

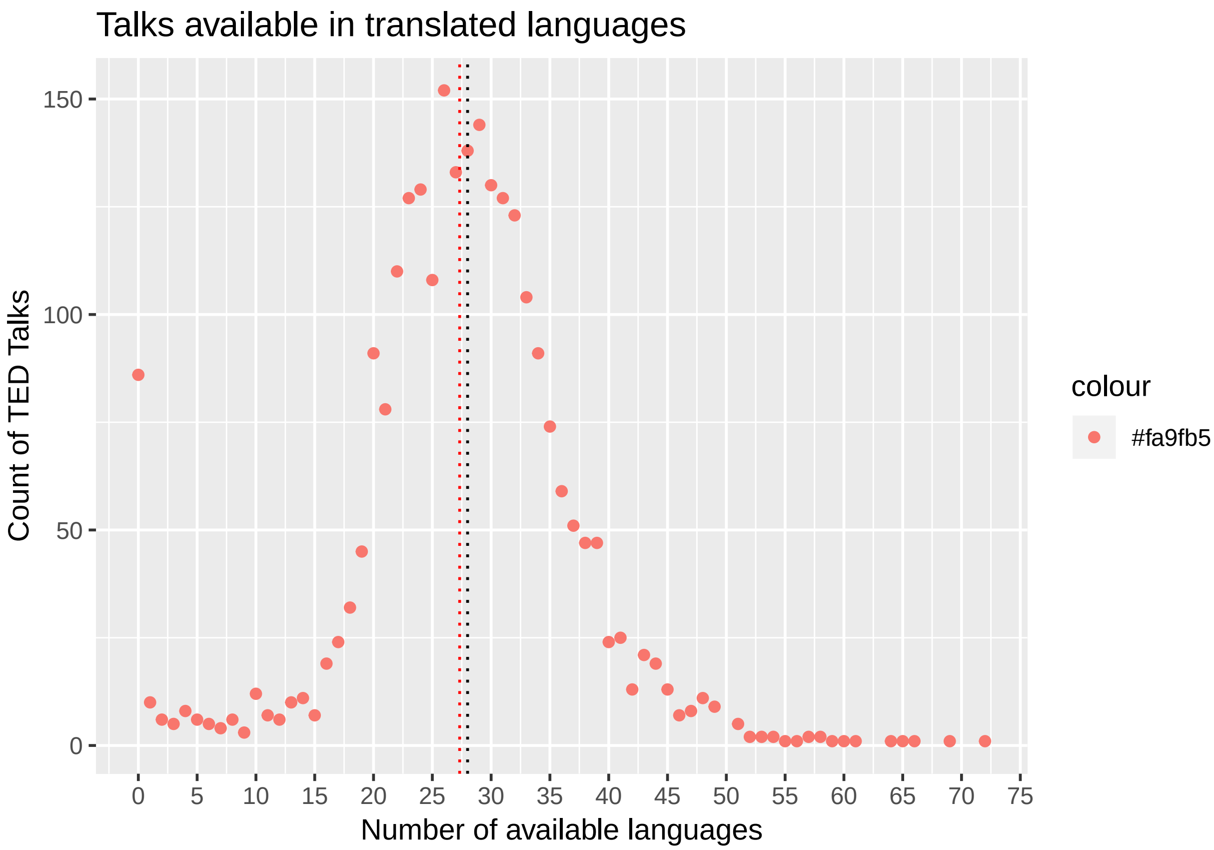


Variable: languages

The variable **languages** is an integer denoting how many languages the talk is available to view in. Presumeably an increase in translations should increase views.

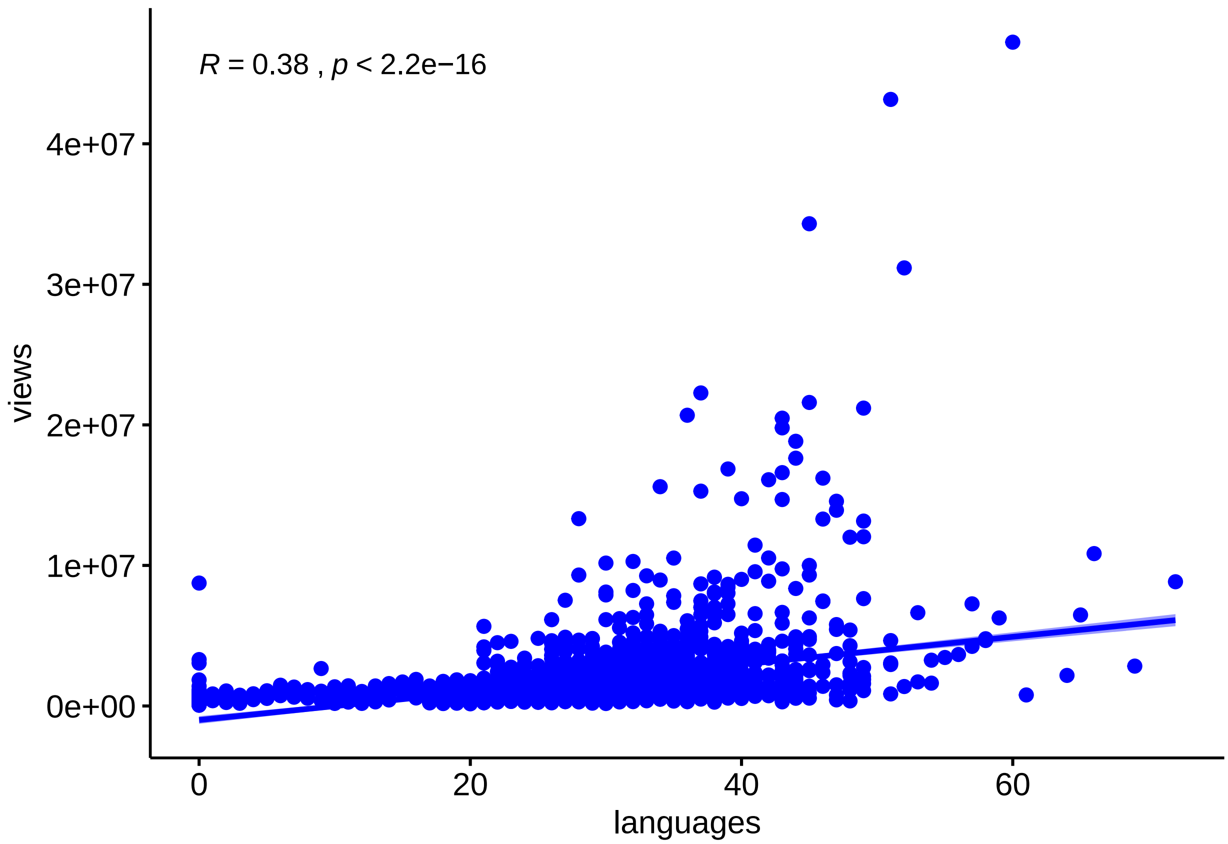
In the graph below we can see that there is a fairly normal distribution for the number of languages.

```
languages <- TedRaw %>%
  group_by(languages) %>%
  tally(name ="numTalks")


#Number of Languages
ggplot(data=languages, aes(x=languages, y=numTalks)) +
  geom_point(stat = "identity", aes(color = '#fa9fb5'))+
  #geom_bar(stat = "identity", fill = '#fa9fb5')+
  geom_vline(aes(xintercept = mean(TedRaw$languages)),col='red',
             size=.5, linetype="dotted")+
  geom_vline(aes(xintercept = median(TedRaw$languages)),col='black',
             size=.5, linetype="dotted")+
  labs(x = "Number of available languages", y = "Count of TED Talks",
       title = "Talks available in translated languages")+
  scale_x_continuous(breaks = pretty(TedRaw$languages, n = 20))
```



Is **languages** correlated with **views**? A little.. Below we can see that the R - value is 0.38 and is statistically significant. We will keep the variable in future models.

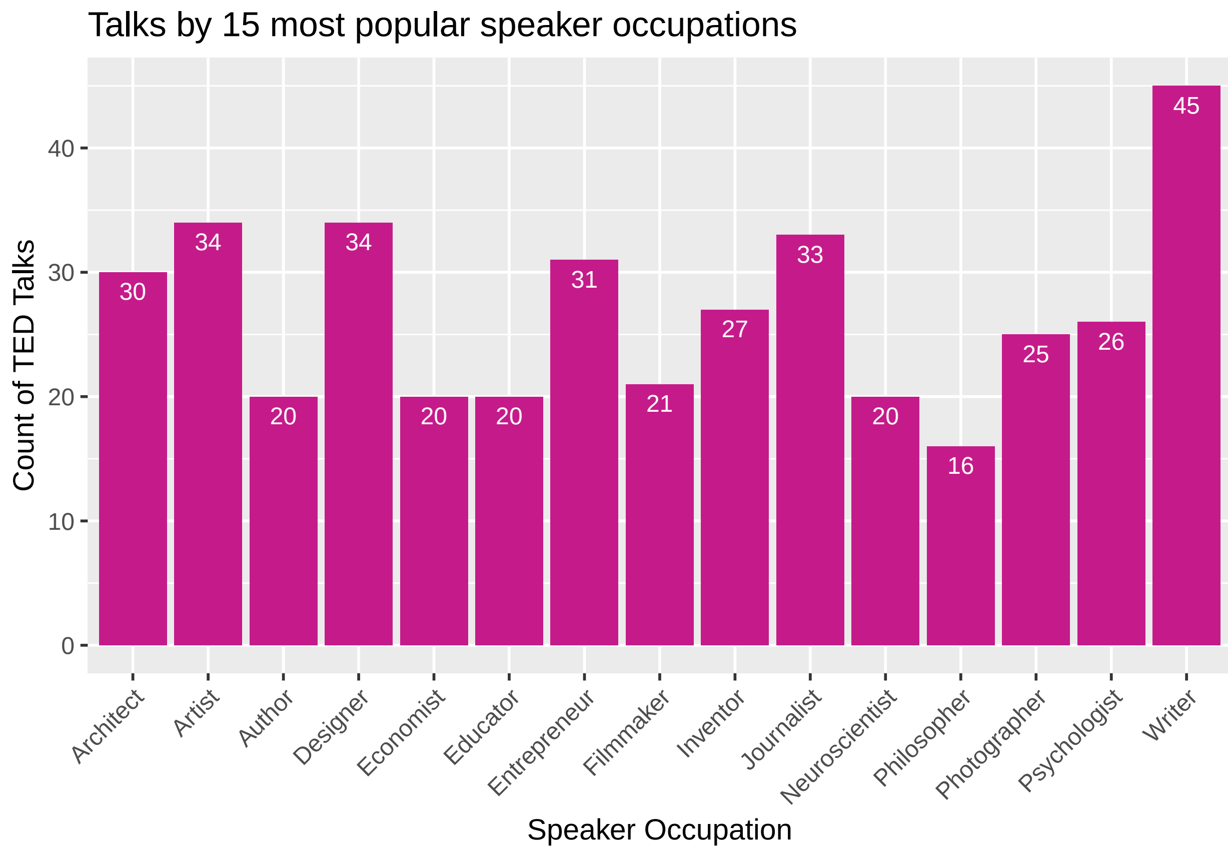
```
ggscatter(TedRaw, x = "languages", y = "views",
          add = "reg.line", conf.int = TRUE, color = "blue",
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "languages", ylab = "views")
```



Variable: speaker_occupation

The variable **speaker_occupation** is the occupation of the main speaker. There are 1459 unique occupations. Below is a graph indicating the top 15 occupations. Since they make up 402 talks (16%) it could be worth exploring breaking the top 15 or 10 occupations into boolean predictors in a future model.

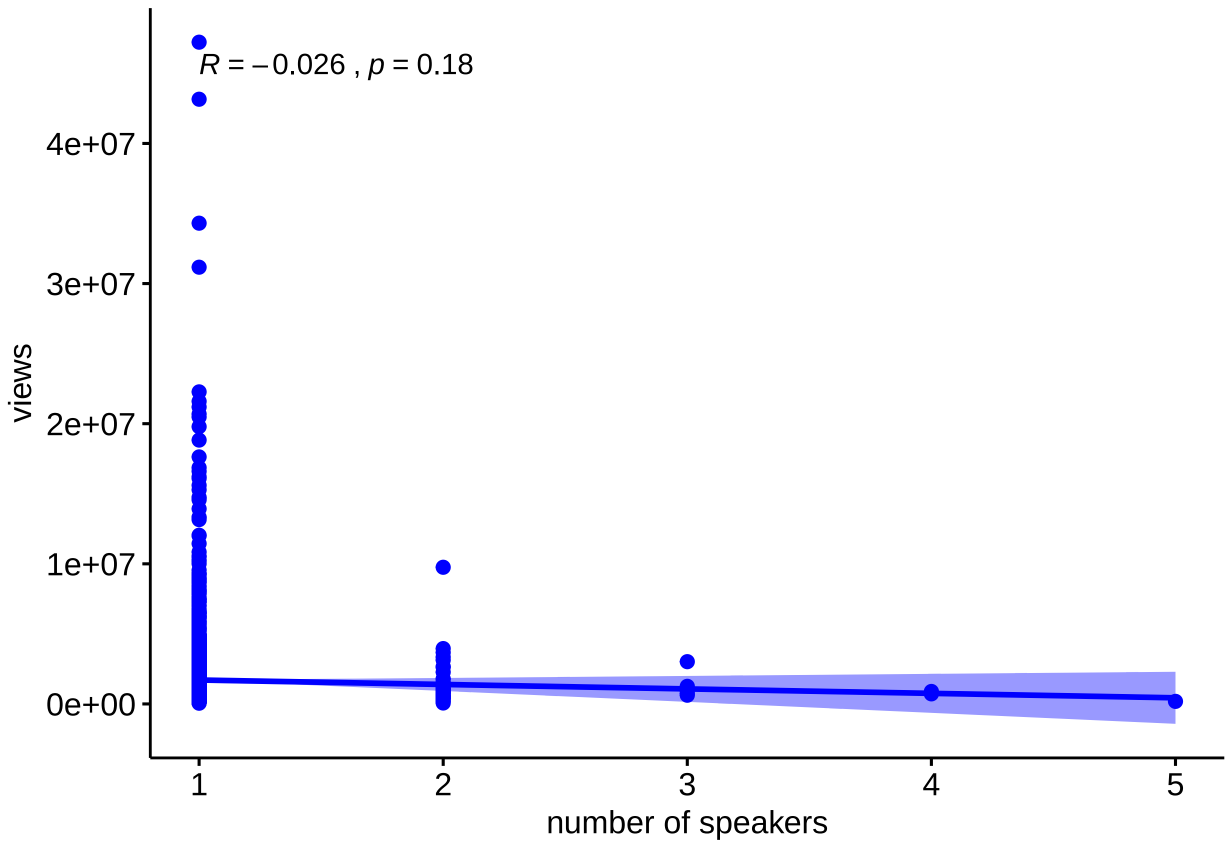
```
##speaker occupation
speaker_occupation <- TedRaw %>%
  group_by(speaker_occupation) %>%
  tally(name ="numTalks")

speaker_occupationtop <- arrange(speaker_occupation, desc(numTalks))
speaker_occupationtop <- head(speaker_occupationtop,15)

ggplot(data=speaker_occupationtop, aes(x=speaker_occupation, y=numTalks)) +
  geom_bar(stat = "identity", fill = '#c51b8a')+
  geom_text(aes(label = numTalks), vjust = 1.6, color = "white", size = 3)+
  labs(x = "Speaker Occupation", y = "Count of TED Talks",
       title = "Talks by 15 most popular speaker occupations")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Variable: num_speaker

The variable **num_speaker** contains the number of speakers in the TED talk.In the graph below we can see the overwhelming majority of talks had 1 speaker. This isn't really an interesting metric and will likely not be included in future models.

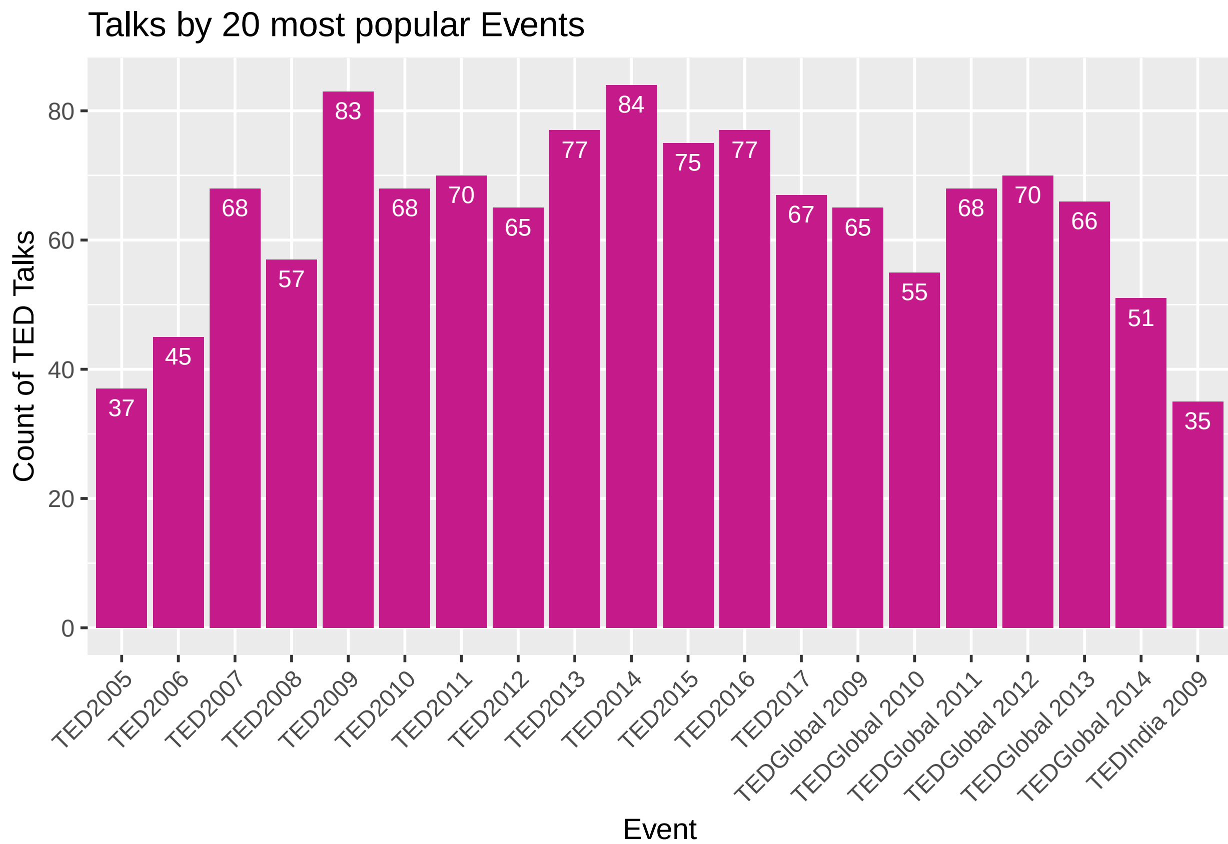
```r
ggscatter(TedRaw, x = "num_speaker", y = "views",
          add = "reg.line", conf.int = TRUE, color = "blue",
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "number of speakers", ylab = "views")
```



Variable: event

The variable **event** is the name of the TED event the TED Talk was given at. There are 355 unique events. Below is a graph showing the top events. Since they make up 1283 (50%) of the talks, there is potential to break this into boolean predictors. Something to be aware of is that the events mostly contain a year in the name so this variable could contain duplicate information as the **filmed_date** variable, so we would likely want to select only one to loop into future models.

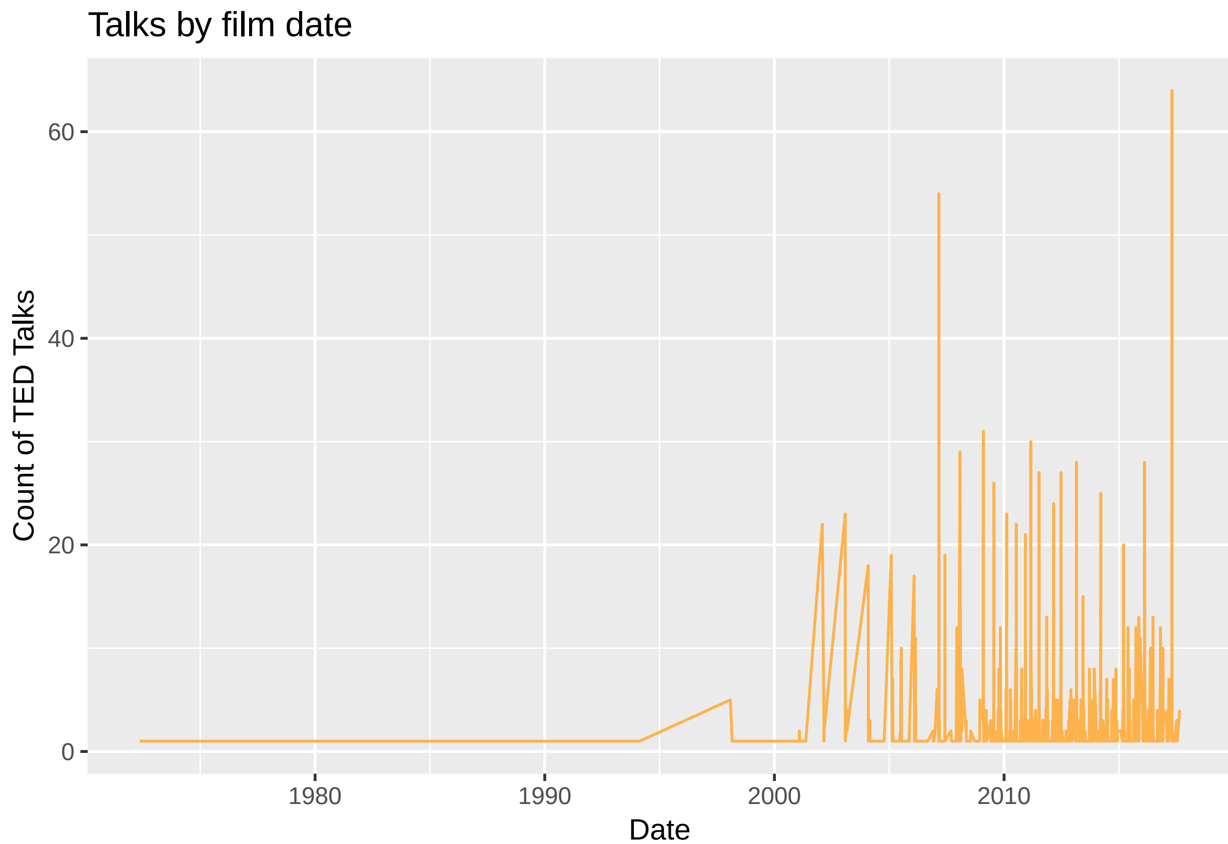
```
event <- TedRaw %>%
  group_by(event) %>%
  tally(name ="numTalks")

eventtop <- arrange(event, desc(numTalks))
eventtop <- head(eventtop,20)
ggplot(data=eventtop, aes(x=event, y=numTalks)) +
  #geom_point(stat = "identity", aes(color = '#2b8cbe'))+
  geom_bar(stat = "identity", fill = '#c51b8a')+
  geom_text(aes(label = numTalks), vjust = 1.6, color = "white", size = 3)+
  labs(x = "Event", y = "Count of TED Talks",
       title = "Talks by 20 most popular Events")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Variable: film_date

The variable **film_date** is the date the TED Talk was filmed. Below is a plot of TED talks by their film date. We likely won't use **film_date** as predictor in favor of **event**.

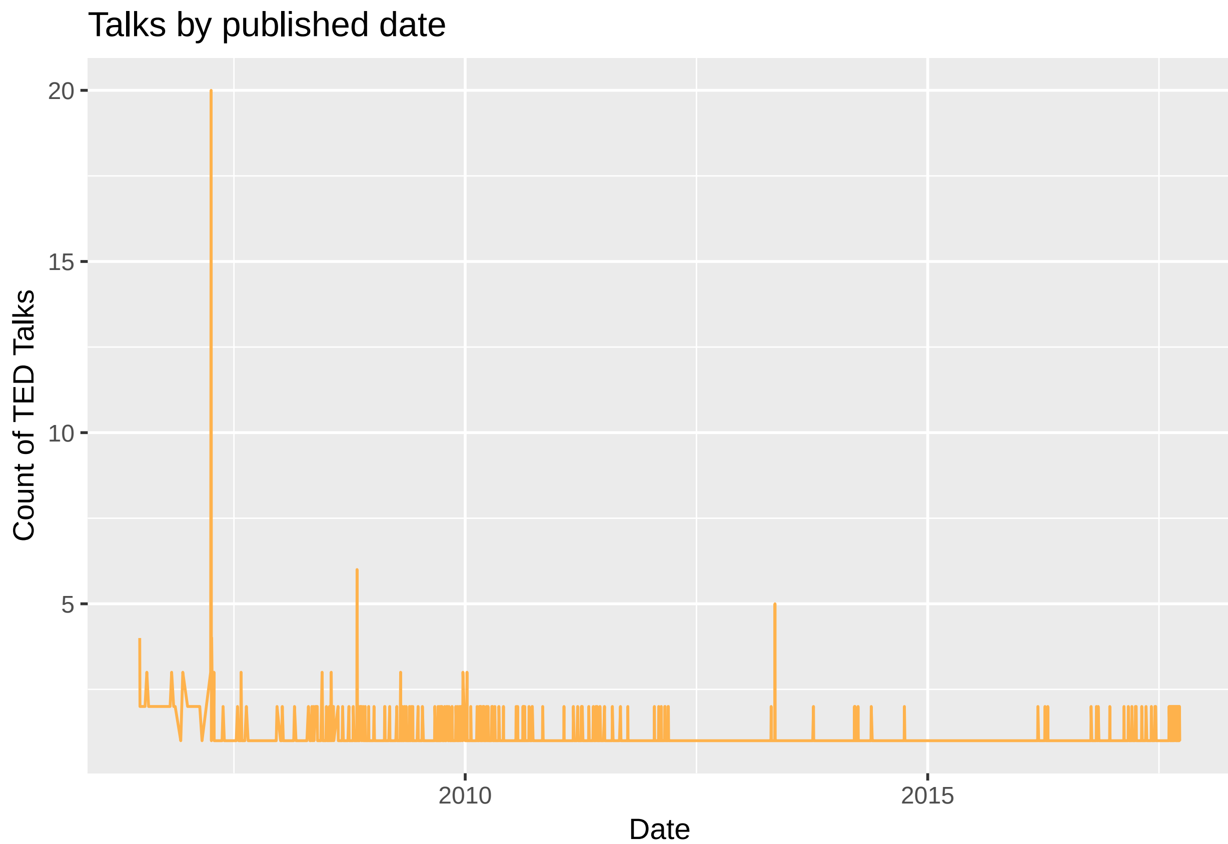
```
#film_date
film_date <- anydate(TedRaw$film_date)
TedRaw$film_date_clean <- film_date


film_date_clean <- TedRaw %>%
  group_by(film_date_clean) %>%
  tally(name ="numTalks")

ggplot(data = film_date_clean, aes(x = film_date_clean, y = numTalks))+
  geom_line(color = "#feb24c")+
  labs(x = "Date", y = "Count of TED Talks",
       title = "Talks by film date")
```



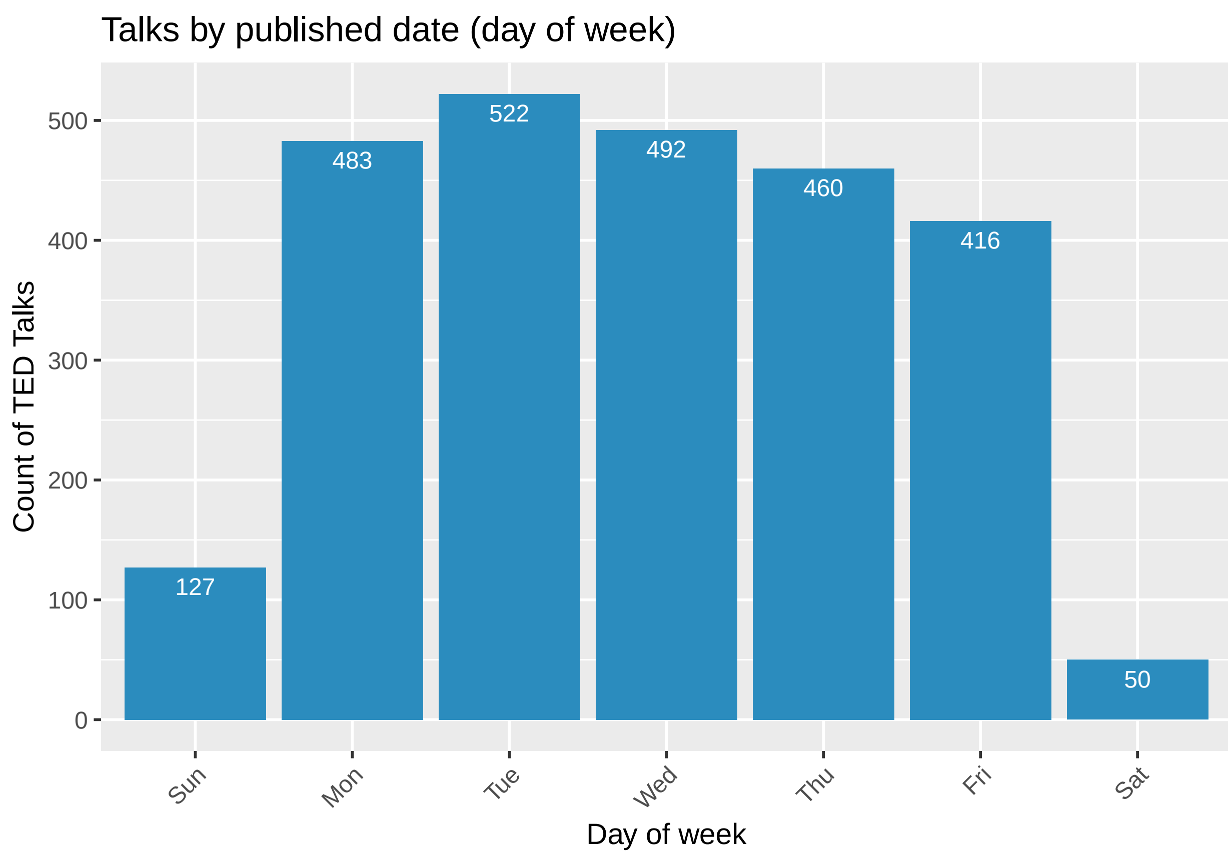
Variable: published_date

The variable **published_date** is the date the TED Talk was published to the TED website. Below are 3 graphs, the first indicating the talks based on their publishing date, then talks based on week day of publish, and talks based on month of publish. The months show no distinct trend. The days of the week indicates the majority of talks are posted on weekdays.

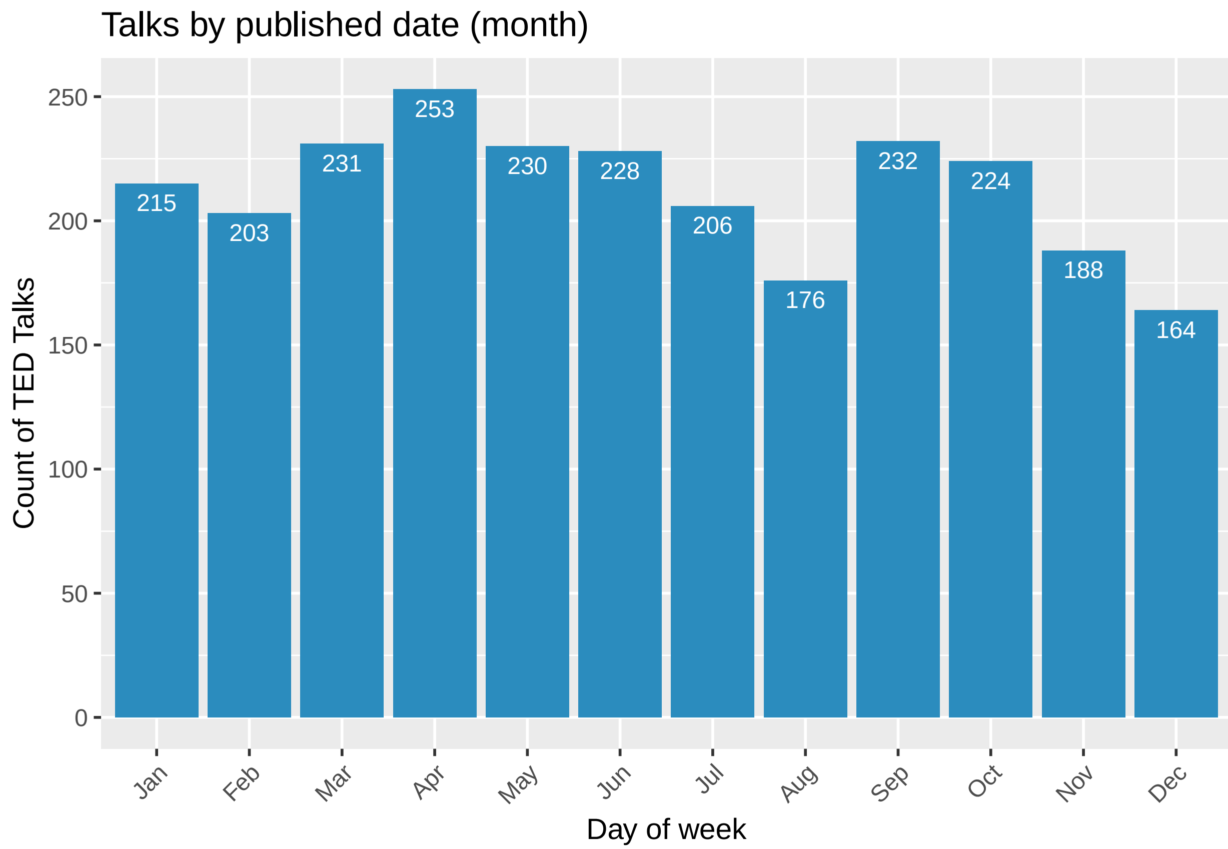
```
#published_date

TedRaw$published_date_clean <- anydate(TedRaw$published_date)

published_date_clean <- TedRaw %>%
  group_by(published_date_clean) %>%
  tally(name ="numTalks")

ggplot(data = published_date_clean, aes(x = published_date_clean, y = numTalks))+
  geom_line(color = "#feb24c")+
  labs(x = "Date", y = "Count of TED Talks",
       title = "Talks by published date")
```




```r
#day of week

TedRaw$published_date_wday <- wday(TedRaw$published_date_clean, label = TRUE)

published_date_wday <- TedRaw %>%
  group_by(published_date_wday) %>%
  tally(name ="numTalks")

ggplot(data=published_date_wday, aes(x=published_date_wday, y=numTalks)) +
  #geom_point(stat = "identity", aes(color = '#2b8cbe'))+
  geom_bar(stat = "identity", fill = '#2b8cbe')+
  geom_text(aes(label = numTalks), vjust = 1.6, color = "white", size = 3)+
  labs(x = "Day of week", y = "Count of TED Talks",
       title = "Talks by published date (day of week)")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

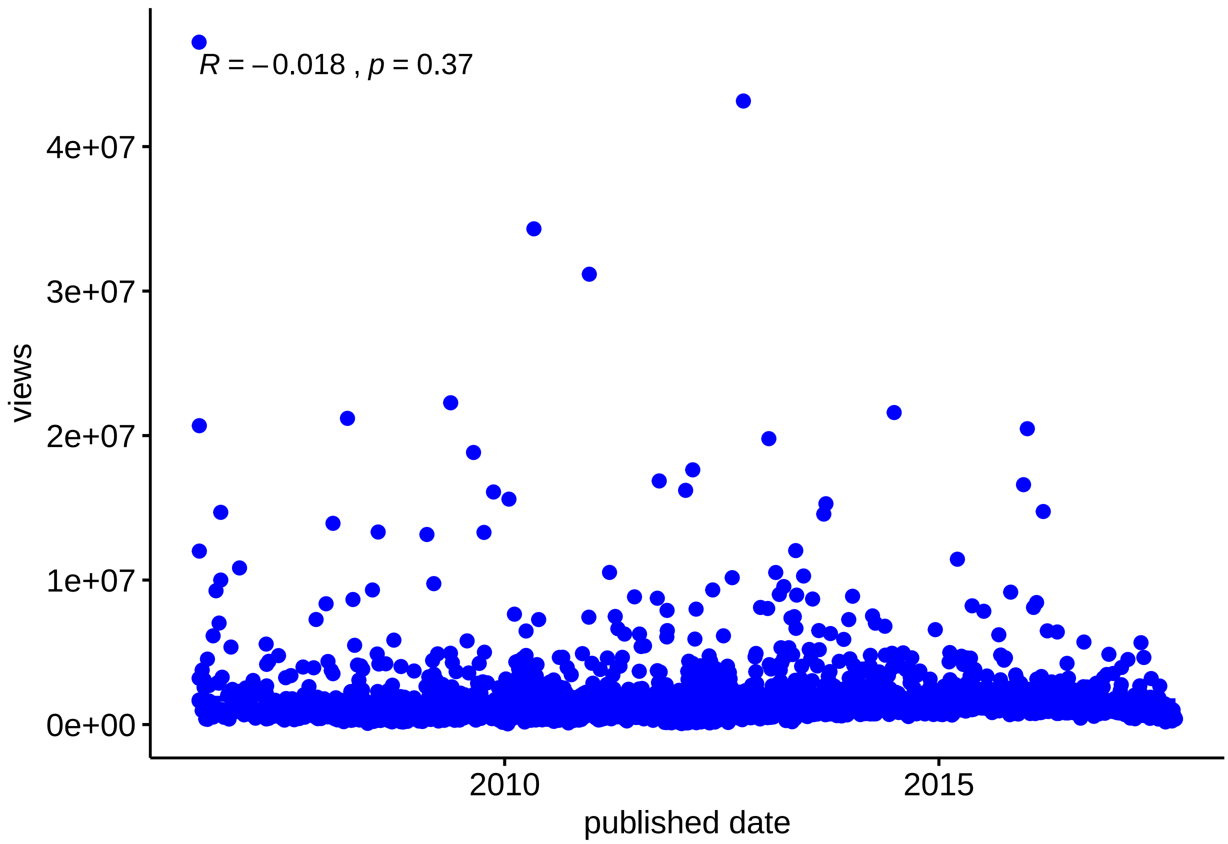



```
#month

TedRaw$published_date_month <- month(TedRaw$published_date_clean, label = TRUE)

published_date_month <- TedRaw %>%
  group_by(published_date_month) %>%
  tally(name ="numTalks")

ggplot(data=published_date_month, aes(x=published_date_month, y=numTalks)) +
  #geom_point(stat = "identity", aes(color = '#2b8cbe'))+
  geom_bar(stat = "identity", fill = '#2b8cbe')+
  geom_text(aes(label = numTalks), vjust = 1.6, color = "white", size = 3)+
  labs(x = "Day of week", y = "Count of TED Talks",
       title = "Talks by published date (month)")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Below we test for correlation and we can see that **published_date** is not correlated with views. In order to simplify we may want to keep **published_date** out of future models.

```
ggscatter(TedRaw, x = "published_date_clean", y = "views",
          add = "reg.line", conf.int = TRUE, color = "blue",
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "published date", ylab = "views")
```



Variable: title

The variable **title** is the official title of the TED Talk. There are 2550 unique titles. Including the full title as a predictor really isn't worthwhile, but there may some common words that we could as as boolean predictors to our model.

Below is a word frequency analysis on the title value. Outputted are the top 20 words. The majority of which would make good boolean predictors for future models.

```
title_corpus <- Corpus(VectorSource(TedRaw$title))

# Convert the text to lower case
title_corpus <- tm_map(title_corpus, content_transformer(tolower))
# Remove numbers
title_corpus <- tm_map(title_corpus, removeNumbers)
# Remove english common stopwords
title_corpus <- tm_map(title_corpus, removeWords, stopwords("english"))
# Remove punctuation
title_corpus <- tm_map(title_corpus, removePunctuation)

dtm_title <- TermDocumentMatrix(title_corpus)
m_title <- as.matrix(dtm_title)
v_title <- sort(rowSums(m_title),decreasing=TRUE)
d_title <- data.frame(word = names(v_title),freq=v_title)
head(d_title, 20)
```

```
##              word freq
```

```
## can         can  102
## life       life   82
## new         new   73
## world     world   69
## future   future   53
## art         art   51
## make       make   45
## design   design   36
## brain     brain   35
## better   better   35
## love       love   33
## change   change   33
## story     story   32
## need       need   32
## science science   30
## power     power   30
## time       time   30
## human     human   30
## women     women   29
## one         one   27
```

Variable: description

The variable **description** is a snippet of text summarizing what the talk is about. There are 2550 unique descriptions. Including the full description as a predictor really isn't feasible, but there may be some common words that we could as as boolean predictors to our model.

Below is a word frequency analysis on the description value. Outputted are the top 50 words. These words feel less like they capture the theme of the talk than the title, but a few might still make decent boolean predictors for future models.

```r
############################################
#Work on finding frequency of the description column
############################################
description_corpus <- Corpus(VectorSource(TedRaw$description))

# Convert the text to lower case
description_corpus <- tm_map(description_corpus, content_transformer(tolower))
# Remove numbers
description_corpus <- tm_map(description_corpus, removeNumbers)
# Remove english common stopwords
description_corpus <- tm_map(description_corpus, removeWords, stopwords("english"))
# Remove punctuation
description_corpus <- tm_map(description_corpus, removePunctuation)

dtm_description <- TermDocumentMatrix(description_corpus)
m_description <- as.matrix(dtm_description)
v_description <- sort(rowSums(m_description),decreasing=TRUE)
d_description <- data.frame(word = names(v_description),freq=v_description)
head(d_description, 50)
```

```
##                 word freq
## talk            talk  699
## can              can  593
## world          world  427
```

```
## new               new  415
## says             says  411
## shares         shares  326
## people         people  324
## ted               ted  296
## shows           shows  282
## one               one  269
## life             life  267
## like             like  250
## make             make  239
## way               way  226
## human           human  207
## work             work  202
## just             just  193
## help             help  184
## story           story  180
## even             even  179
## time             time  169
## years           years  164
## makes           makes  153
## talks           talks  148
## will             will  148
## data             data  142
## future         future  142
## change         change  140
## powerful     powerful  140
## now               now  136
## know             know  135
## two               two  130
## using           using  130
## science       science  130
## asks             asks  129
## stories       stories  128
## many             many  128
## think           think  126
## tells           tells  125
## learn           learn  124
## -                   -  124
## look             look  123
## technology technology  121
## need             need  120
## fellow         fellow  119
## first           first  118
## lives           lives  116
## get               get  114
## might           might  114
## design         design  113
```

Variable: tags

The variable **tags** is a set of phrases that describe that talk. Each talk has a few tags associated with. Below is a word frequency analysis on the top 30 tags. These would likely make great boolean predictors since they appear fairly frequently amongst the dataset.

```r
TedRaw$tags <- TedRaw$tags %>%
  str_replace_all('\\[','') %>%
  str_replace_all('\\]','')   %>%
  str_replace_all("\\'",' ') %>%
  str_replace_all(',',' ') %>%
  tolower()

#talk_tags <- unnest_tokens(ted3,tags1,tags) %>% select(sno,tags1)
#datatable(head(talk_tags,10))

tags_corpus <- Corpus(VectorSource(TedRaw$tags))

# Remove punctuation
tags_corpus <- tm_map(tags_corpus, removePunctuation)
# Remove numbers
tags_corpus <- tm_map(tags_corpus, removeNumbers)
# Remove english common stopwords
tags_corpus <- tm_map(tags_corpus, removeWords, stopwords("english"))


dtm_tags <- TermDocumentMatrix(tags_corpus)
m_tags <- as.matrix(dtm_tags)
v_tags <- sort(rowSums(m_tags),decreasing=TRUE)
d_tags <- data.frame(word = names(v_tags),freq=v_tags)
head(d_tags, 30)
```

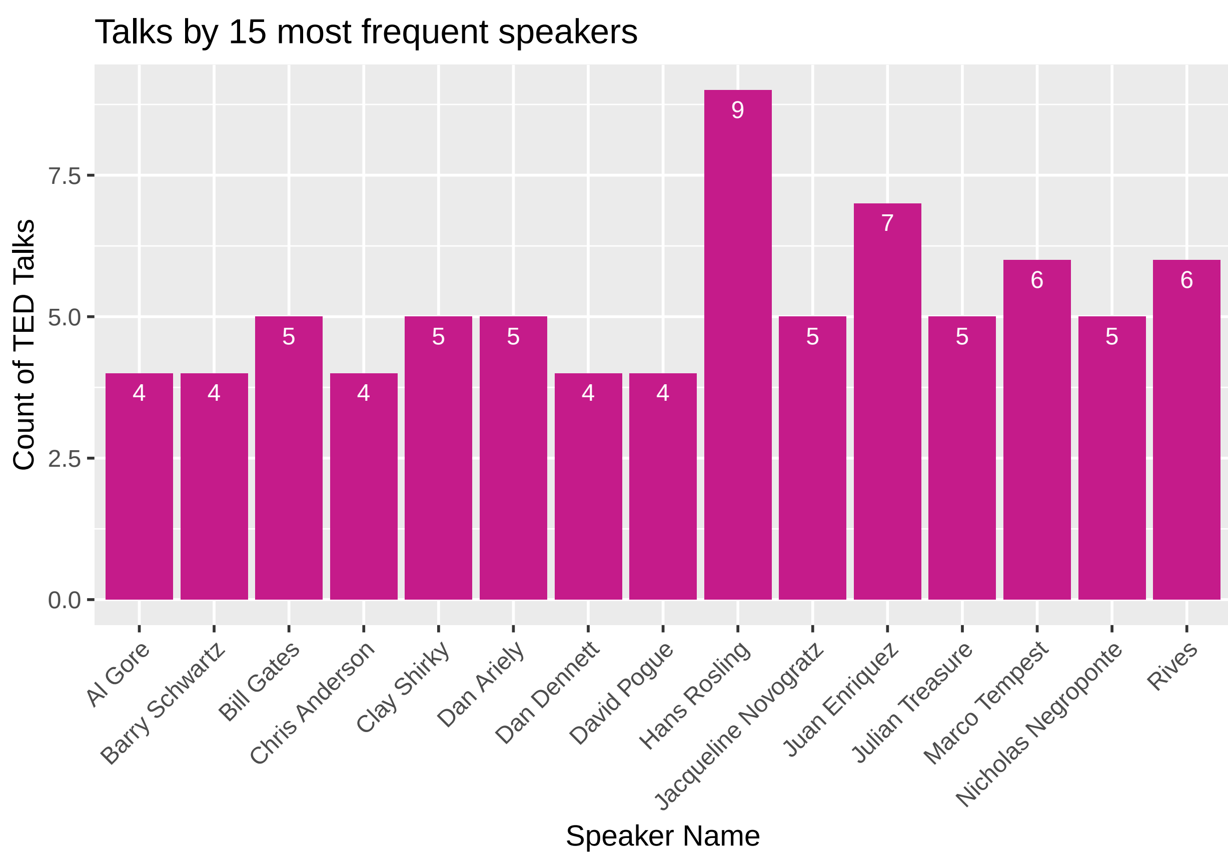
```
##                        word freq
## technology       technology  727
## science             science  675
## global               global  565
## design               design  526
## issues               issues  501
## health               health  489
## culture             culture  486
## tedx                   tedx  450
## business           business  374
## change               change  305
## entertainment entertainment  299
## art                     art  289
## social               social  270
## ted                     ted  254
## biology             biology  234
## innovation       innovation  229
## society             society  224
## music                 music  220
## brain                 brain  207
## future               future  195
## communication communication  191
## creativity       creativity  189
## economics         economics  187
## humanity           humanity  182
## collaboration collaboration  174
## environment     environment  165
## medicine           medicine  162
```

```
## activism            activism  157
## education          education  153
## community          community  148
```

Variable: main_speaker

The variable **main_speaker** contains the name of the speaker. There are 2156 unique speakers of the 2550 TED talks. Below is a graph showing the top 15 speakers by number of TED talks. This could be made a boolean predictor in a future model.

```
main_speaker <- TedRaw %>%
  group_by(main_speaker) %>%
  tally(name ="numTalks")

main_speaker <- arrange(main_speaker, desc(numTalks))
main_speakertop <- head(main_speaker,15)

ggplot(data=main_speakertop, aes(x=main_speaker, y=numTalks)) +
  #geom_point(stat = "identity", aes(color = '#2b8cbe'))+
  geom_bar(stat = "identity", fill = '#c51b8a')+
  geom_text(aes(label = numTalks), vjust = 1.6, color = "white", size = 3)+
  labs(x = "Speaker Name", y = "Count of TED Talks",
       title = "Talks by 15 most frequent speakers")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Variables not included

The following variables were removed from the dataset.
name – this contains duplicate information as it is the title and speaker combined.
ratings – too complex to parse left out for simplification.
related talks – this is a list of urls, it provides no meaning.
url – the url of the talk provides no meaning.

Data Transformation

Below is the code to tranform the data based on our findings in the EDA section. The following variables were removed: film_date, name, num_speaker, published_date, ratings, and related_talks. **title** was transformed to 20 boolean columns based on if the title contains the top 20 words. **tags** was transformed to 50 boolean columns based on if the talk contains some of the top 50 tags. **speaker_occupation** was transformed to 15 boolean columns representing if the occupation contains the top 15 occupations- note to catch more talks if the occupation was a part of another occupation it was considered (ex. singer/song writer is considered a writer).**main_speaker** was transformed to 15 boolean colums denoting if the talk was given by one of the 15 most frequent speakers. **event** was transformed to 15 boolean columsn denoting if the talk happened at that event. **description** was transformed to 30 boolean columsn denoting if the talk description contains that word.

```
library(tidyverse)

TedRaw <- read.csv(file="ted_main.csv", header=TRUE, sep=",")

#comments(as is)
#duration (as is)
#languages (as is)
#views (as is)

TedRaw$film_date <- NULL
TedRaw$name <- NULL
TedRaw$num_speaker <- NULL
TedRaw$published_date <- NULL
TedRaw$ratings <- NULL
TedRaw$related_talks <- NULL
TedRaw$url <- NULL

####################
## title
##loop through top 20 words and create boolean columns for them

for(i in 1:20){
  col_name <- paste("tags",d_title$word[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(d_title$word[i], TedRaw$title, fixed = TRUE))

}

TedRaw$title <- NULL
```

```
####################
## tags
##loop through top 50 tags and create boolean columns for them

for(i in 1:50){
  col_name <- paste("tags",d_tags$word[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(d_tags$word[i], TedRaw$tags, fixed = TRUE))

}

TedRaw$tags <- NULL



####################
## for speaker_occupation
##loop through top 15 occupations and create boolean columns for them

for(i in 1:15){
  col_name <- paste("speaker_occupation",speaker_occupationtop$speaker_occupation[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(speaker_occupationtop$speaker_occupation[i], TedRaw$speaker_occupation, fixed = TRUE
    )

}

TedRaw$speaker_occupation <- NULL


####################
## for main_speaker
##loop through top 15 speakers and create boolean columns for them

for(i in 1:15){
  col_name <- paste("speaker",main_speakertop$main_speaker[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(main_speakertop$main_speaker[i], TedRaw$main_speaker, fixed = TRUE))

}

TedRaw$main_speaker <- NULL


####################
# for event
## loop through the top 20 events and create boolean columns for them
```

```r
for(i in 1:20){
  col_name <- paste("event",eventtop$event[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(eventtop$event[i], TedRaw$event, fixed = TRUE))

}

TedRaw$event <- NULL



####################
## for description
##loop through top 30 description words and create boolean columns for them
for(i in 1:30){
  col_name <- paste("description",d_description$word[i], sep = "_")

  TedRaw <-  TedRaw %>%
    mutate(!!col_name :=
             grepl(d_description$word[i], TedRaw$description, fixed = TRUE))

}

TedRaw$description <- NULL

TedClean <- TedRaw %>% dplyr::rename_all(funs(make.names(.)))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##
## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

Building and evaluating Models

In order to understand what predictors influence the number of views the most we built a few different models.

Linear Regression

Below are the results for the full linear regression model. For the full linear regression model we can see that the following predictors were deemed as significant to the model and contributed to an increase in views: comments, duration, languages, tag contains “life”, and tag contains “make”. The following predictors were deemed as significant to the model and contributed to a decrease in views: tag contains “can”, tag contains “new”, tag contains “world”, tag contains “art”, and tag contains “design”. The R-squared for the linear model is 0.4359, which is pretty low. Ideally, we would improve on this with other models.

```r
##Full linear regression model
set.seed(1)
summary(lm(views~., TedClean))
```

```
##
## Call:
## lm(formula = views ~ ., data = TedClean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -23884807   -694789   -153029    428753  27986856
##
## Coefficients: (2 not defined because of singularities)
##                                   Estimate Std. Error t value
## (Intercept)                     -1538413.3   220818.3  -6.967
## comments                            4019.4      157.8  25.472
## duration                             590.2      119.1   4.954
## languages                          79765.0     5228.2  15.257
## tags_canTRUE                      -36136.4   190826.7  -0.189
## tags_lifeTRUE                     195846.2   190753.8   1.027
## tags_newTRUE                     -185762.3   248207.5  -0.748
## tags_worldTRUE                    -12342.4   225605.6  -0.055
## tags_futureTRUE                   -17713.9   173329.0  -0.102
## tags_artTRUE                     -115305.2   141876.8  -0.813
## tags_makeTRUE                     332649.7   256188.9   1.298
## tags_designTRUE                  -116635.8   127718.7  -0.913
## tags_brainTRUE                    557181.7   190375.1   2.927
## tags_betterTRUE                   124867.3   352024.2   0.355
## tags_loveTRUE                     333601.2   356548.5   0.936
## tags_changeTRUE                  -209282.3   232737.2  -0.899
## tags_storyTRUE                    188771.8   298191.2   0.633
## tags_needTRUE                    -226007.8   301842.6  -0.749
## tags_scienceTRUE                 -305817.3   118083.2  -2.590
## tags_powerTRUE                   1184628.3   310324.9   3.817
## tags_timeTRUE                    -325273.7   322533.4  -1.008
## tags_humanTRUE                   -169519.7   302179.0  -0.561
## tags_womenTRUE                   -283739.3   185977.1  -1.526
## tags_oneTRUE                      -92088.4   233465.7  -0.394
## tags_technologyTRUE               -88003.5   101649.5  -0.866
## tags_globalTRUE                  -186290.1   451846.8  -0.412
## tags_issuesTRUE                  -479490.0   457398.2  -1.048
## tags_healthTRUE                   199786.9   162958.5   1.226
## tags_cultureTRUE                 -158963.3   103656.9  -1.534
## tags_tedxTRUE                           NA         NA      NA
## tags_businessTRUE                 373479.5   126352.8   2.956
## tags_entertainmentTRUE            249247.1   139808.3   1.783
## tags_socialTRUE                   201372.3   250724.3   0.803
## tags_tedTRUE                      210376.3   362234.5   0.581
## tags_biologyTRUE                  -51532.2   167041.7  -0.308
## tags_innovationTRUE               153602.8   163197.5   0.941
## tags_societyTRUE                  269379.1   171684.5   1.569
## tags_musicTRUE                   -293814.0   207254.9  -1.418
## tags_communicationTRUE            283441.2   162292.5   1.746
## tags_creativityTRUE               188243.0   157664.1   1.194
```

```
## tags_economicsTRUE                       -403260.9   176271.8  -2.288
## tags_humanityTRUE                         270229.4   178956.6   1.510
## tags_collaborationTRUE                     -8006.5   168122.5  -0.048
## tags_environmentTRUE                      -59201.4   190362.3  -0.311
## tags_medicineTRUE                        -135391.9   201359.2  -0.672
## tags_activismTRUE                        -144970.9   173781.4  -0.834
## tags_educationTRUE                        137680.9   186634.5   0.738
## tags_communityTRUE                       -142010.7   188579.5  -0.753
## tags_historyTRUE                          -13939.6   178178.0  -0.078
## tags_childrenTRUE                        -151652.0   183934.6  -0.824
## tags_fellowsTRUE                                NA         NA      NA
## tags_performanceTRUE                      470169.9   216858.4   2.168
## tags_inventionTRUE                         -1434.2   199749.1  -0.007
## tags_psychologyTRUE                       997649.3   210586.4   4.737
## tags_careTRUE                            -425112.5   237788.9  -1.788
## tags_politicsTRUE                        -490388.4   193068.4  -2.540
## tags_citiesTRUE                          -267501.1   196893.1  -1.359
## tags_energyTRUE                          -364944.4   246716.6  -1.479
## tags_mediaTRUE                           -180182.1   213717.6  -0.843
## tags_storytellingTRUE                    -225184.7   197196.2  -1.142
## tags_natureTRUE                           405821.5   210543.5   1.927
## tags_warTRUE                              -33991.2   169478.6  -0.201
## tags_identityTRUE                          -4780.2   217783.5  -0.022
## tags_computersTRUE                       -106266.6   203583.8  -0.522
## tags_engineeringTRUE                      -18921.4   211811.0  -0.089
## tags_animalsTRUE                           31039.6   211147.9   0.147
## speaker_occupation_WriterTRUE             379693.0   241025.6   1.575
## speaker_occupation_ArtistTRUE            -294132.1   282708.7  -1.040
## speaker_occupation_DesignerTRUE          -191238.0   284950.1  -0.671
## speaker_occupation_JournalistTRUE        -307516.9   306602.1  -1.003
## speaker_occupation_EntrepreneurTRUE      -319539.5   314492.5  -1.016
## speaker_occupation_ArchitectTRUE          -88619.5   337019.6  -0.263
## speaker_occupation_InventorTRUE          -130426.3   353497.1  -0.369
## speaker_occupation_PsychologistTRUE         8506.4   414864.0   0.021
## speaker_occupation_PhotographerTRUE      -441688.3   381471.9  -1.158
## speaker_occupation_FilmmakerTRUE         -250320.1   428524.9  -0.584
## speaker_occupation_AuthorTRUE             251018.3   291792.4   0.860
## speaker_occupation_EconomistTRUE           22161.4   395589.7   0.056
## speaker_occupation_EducatorTRUE          -187584.1   401950.9  -0.467
## speaker_occupation_NeuroscientistTRUE  -798468.3   370313.7  -2.156
## speaker_occupation_PhilosopherTRUE    -1025905.5   434133.0  -2.363
## speaker_Hans.RoslingTRUE                  295101.4   683764.6   0.432
## speaker_Juan.EnriquezTRUE                 -57380.6   761539.9  -0.075
## speaker_Marco.TempestTRUE                 347718.8   816261.2   0.426
## speaker_RivesTRUE                        -701793.0   813376.0  -0.863
## speaker_Bill.GatesTRUE                   -296840.4   887479.5  -0.334
## speaker_Clay.ShirkyTRUE                 -1182501.0   904533.7  -1.307
## speaker_Dan.ArielyTRUE                      8420.3   900843.4   0.009
## speaker_Jacqueline.NovogratzTRUE          -23349.8   895235.5  -0.026
## speaker_Julian.TreasureTRUE              3705602.9   892848.6   4.150
## speaker_Nicholas.NegroponteTRUE           -95707.7   897390.4  -0.107
## speaker_Al.GoreTRUE                     -1933730.9  1024983.2  -1.887
## speaker_Barry.SchwartzTRUE                -64809.7  1057497.7  -0.061
## speaker_Chris.AndersonTRUE               -399583.8   882785.4  -0.453
```

```
## speaker_Dan.DennettTRUE                -782086.1  1070761.7  -0.730
## speaker_David.PogueTRUE                 -27938.2   987877.9  -0.028
## event_TED2014TRUE                        70879.1   226189.9   0.313
## event_TED2009TRUE                      -271310.0   229601.8  -1.182
## event_TED2013TRUE                      -265637.9   235926.0  -1.126
## event_TED2016TRUE                       481461.5   246197.1   1.956
## event_TED2015TRUE                       212215.7   239986.6   0.884
## event_TED2011TRUE                      -534137.3   250038.5  -2.136
## event_TEDGlobal.2012TRUE                241918.8   247259.4   0.978
## event_TED2007TRUE                       -43218.7   254387.1  -0.170
## event_TED2010TRUE                     -1137396.4   250593.0  -4.539
## event_TEDGlobal.2011TRUE               -761675.1   253417.3  -3.006
## event_TED2017TRUE                      1168879.1   264737.1   4.415
## event_TEDGlobal.2013TRUE                271442.3   252556.1   1.075
## event_TED2012TRUE                      -155068.4   257567.7  -0.602
## event_TEDGlobal.2009TRUE                -20543.5   255782.2  -0.080
## event_TED2008TRUE                        -5725.6   273744.2  -0.021
## event_TEDGlobal.2010TRUE              -1388707.8   274171.9  -5.065
## event_TEDGlobal.2014TRUE                -73612.4   285713.6  -0.258
## event_TED2006TRUE                      1242433.1   310442.8   4.002
## event_TED2005TRUE                       146680.0   336506.4   0.436
## event_TEDIndia.2009TRUE                -683435.6   339766.5  -2.011
## description_talkTRUE                    125864.8    93733.7   1.343
## description_canTRUE                      18793.1    93587.7   0.201
## description_worldTRUE                    34665.9   108145.0   0.321
## description_newTRUE                    -142138.0   114164.5  -1.245
## description_saysTRUE                   -206415.1   116516.0  -1.772
## description_sharesTRUE                 -109690.6   120661.3  -0.909
## description_peopleTRUE                   45538.4   131189.2   0.347
## description_tedTRUE                      52310.2    96800.2   0.540
## description_showsTRUE                   -72367.5   129752.6  -0.558
## description_oneTRUE                      94394.3    94548.2   0.998
## description_lifeTRUE                    -98939.2   123316.0  -0.802
## description_likeTRUE                    104540.8   128815.7   0.812
## description_makeTRUE                    146497.1   140339.4   1.044
## description_wayTRUE                    -155008.3   111861.3  -1.386
## description_humanTRUE                  -158591.7   129959.7  -1.220
## description_workTRUE                    114742.3   114098.6   1.006
## description_justTRUE                     20620.7   142554.0   0.145
## description_helpTRUE                    173578.4   135727.0   1.279
## description_storyTRUE                   -77658.1   138618.6  -0.560
## description_evenTRUE                    119740.1   134927.8   0.887
## description_timeTRUE                    111622.6   132648.1   0.841
## description_yearsTRUE                   -32714.7   166236.7  -0.197
## description_makesTRUE                  -348379.4   209764.5  -1.661
## description_talksTRUE                  -487039.5   189428.2  -2.571
## description_willTRUE                   -101978.2   182313.1  -0.559
## description_dataTRUE                     31542.9   192593.1   0.164
## description_futureTRUE                  -49567.4   181994.8  -0.272
## description_changeTRUE                  202673.2   168000.2   1.206
## description_powerfulTRUE                 13493.4   173167.8   0.078
## description_nowTRUE                      82584.2   122066.9   0.677
##                                       Pr(>|t|)
## (Intercept)                           4.17e-12 ***
```

```
## comments                              < 2e-16 ***
## duration                             7.79e-07 ***
## languages                             < 2e-16 ***
## tags_canTRUE                         0.849821
## tags_lifeTRUE                        0.304667
## tags_newTRUE                         0.454283
## tags_worldTRUE                       0.956376
## tags_futureTRUE                      0.918608
## tags_artTRUE                         0.416463
## tags_makeTRUE                        0.194256
## tags_designTRUE                      0.361216
## tags_brainTRUE                       0.003457 **
## tags_betterTRUE                      0.722836
## tags_loveTRUE                        0.349552
## tags_changeTRUE                      0.368625
## tags_storyTRUE                       0.526757
## tags_needTRUE                        0.454075
## tags_scienceTRUE                     0.009660 **
## tags_powerTRUE                       0.000138 ***
## tags_timeTRUE                        0.313318
## tags_humanTRUE                       0.574856
## tags_womenTRUE                       0.127224
## tags_oneTRUE                         0.693290
## tags_technologyTRUE                  0.386711
## tags_globalTRUE                      0.680167
## tags_issuesTRUE                      0.294606
## tags_healthTRUE                      0.220319
## tags_cultureTRUE                     0.125271
## tags_tedxTRUE                              NA
## tags_businessTRUE                    0.003148 **
## tags_entertainmentTRUE               0.074749 .
## tags_socialTRUE                      0.421960
## tags_tedTRUE                         0.561447
## tags_biologyTRUE                     0.757730
## tags_innovationTRUE                  0.346693
## tags_societyTRUE                     0.116771
## tags_musicTRUE                       0.156424
## tags_communicationTRUE               0.080855 .
## tags_creativityTRUE                  0.232615
## tags_economicsTRUE                   0.022240 *
## tags_humanityTRUE                    0.131168
## tags_collaborationTRUE               0.962021
## tags_environmentTRUE                 0.755833
## tags_medicineTRUE                    0.501400
## tags_activismTRUE                    0.404243
## tags_educationTRUE                   0.460766
## tags_communityTRUE                   0.451491
## tags_historyTRUE                     0.937648
## tags_childrenTRUE                    0.409744
## tags_fellowsTRUE                           NA
## tags_performanceTRUE                 0.030249 *
## tags_inventionTRUE                   0.994272
## tags_psychologyTRUE                  2.29e-06 ***
## tags_careTRUE                        0.073939 .
```

```
## tags_politicsTRUE                        0.011149 *
## tags_citiesTRUE                          0.174397
## tags_energyTRUE                          0.139216
## tags_mediaTRUE                           0.399265
## tags_storytellingTRUE                    0.253596
## tags_natureTRUE                          0.054036 .
## tags_warTRUE                             0.841057
## tags_identityTRUE                        0.982490
## tags_computersTRUE                       0.601733
## tags_engineeringTRUE                     0.928826
## tags_animalsTRUE                         0.883141
## speaker_occupation_WriterTRUE            0.115314
## speaker_occupation_ArtistTRUE            0.298255
## speaker_occupation_DesignerTRUE          0.502203
## speaker_occupation_JournalistTRUE        0.315970
## speaker_occupation_EntrepreneurTRUE      0.309709
## speaker_occupation_ArchitectTRUE         0.792611
## speaker_occupation_InventorTRUE          0.712190
## speaker_occupation_PsychologistTRUE      0.983643
## speaker_occupation_PhotographerTRUE      0.247039
## speaker_occupation_FilmmakerTRUE         0.559178
## speaker_occupation_AuthorTRUE            0.389730
## speaker_occupation_EconomistTRUE         0.955330
## speaker_occupation_EducatorTRUE          0.640768
## speaker_occupation_NeuroscientistTRUE 0.031167 *
## speaker_occupation_PhilosopherTRUE       0.018201 *
## speaker_Hans.RoslingTRUE                 0.666083
## speaker_Juan.EnriquezTRUE                0.939944
## speaker_Marco.TempestTRUE                0.670153
## speaker_RivesTRUE                        0.388325
## speaker_Bill.GatesTRUE                   0.738050
## speaker_Clay.ShirkyTRUE                  0.191234
## speaker_Dan.ArielyTRUE                   0.992543
## speaker_Jacqueline.NovogratzTRUE         0.979194
## speaker_Julian.TreasureTRUE              3.44e-05 ***
## speaker_Nicholas.NegroponteTRUE          0.915075
## speaker_Al.GoreTRUE                      0.059335 .
## speaker_Barry.SchwartzTRUE               0.951137
## speaker_Chris.AndersonTRUE               0.650849
## speaker_Dan.DennettTRUE                  0.465216
## speaker_David.PogueTRUE                  0.977440
## event_TED2014TRUE                        0.754034
## event_TED2009TRUE                        0.237460
## event_TED2013TRUE                        0.260304
## event_TED2016TRUE                        0.050629 .
## event_TED2015TRUE                        0.376633
## event_TED2011TRUE                        0.032762 *
## event_TEDGlobal.2012TRUE                 0.327975
## event_TED2007TRUE                        0.865108
## event_TED2010TRUE                        5.94e-06 ***
## event_TEDGlobal.2011TRUE                 0.002678 **
## event_TED2017TRUE                        1.05e-05 ***
## event_TEDGlobal.2013TRUE                 0.282581
## event_TED2012TRUE                        0.547198
```

```
## event_TEDGlobal.2009TRUE              0.935992
## event_TED2008TRUE                    0.983314
## event_TEDGlobal.2010TRUE             4.39e-07 ***
## event_TEDGlobal.2014TRUE             0.796704
## event_TED2006TRUE                    6.47e-05 ***
## event_TED2005TRUE                    0.662955
## event_TEDIndia.2009TRUE              0.044385 *
## description_talkTRUE                 0.179467
## description_canTRUE                  0.840866
## description_worldTRUE                0.748579
## description_newTRUE                  0.213243
## description_saysTRUE                 0.076594 .
## description_sharesTRUE               0.363399
## description_peopleTRUE               0.728532
## description_tedTRUE                  0.588976
## description_showsTRUE                0.577078
## description_oneTRUE                  0.318199
## description_lifeTRUE                 0.422446
## description_likeTRUE                 0.417128
## description_makeTRUE                 0.296647
## description_wayTRUE                  0.165961
## description_humanTRUE                0.222465
## description_workTRUE                 0.314689
## description_justTRUE                 0.884998
## description_helpTRUE                 0.201063
## description_storyTRUE                0.575376
## description_evenTRUE                 0.374931
## description_timeTRUE                 0.400155
## description_yearsTRUE                0.844004
## description_makesTRUE                0.096882 .
## description_talksTRUE                0.010197 *
## description_willTRUE                 0.575970
## description_dataTRUE                 0.869918
## description_futureTRUE               0.785372
## description_changeTRUE               0.227787
## description_powerfulTRUE             0.937897
## description_nowTRUE                  0.498757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1931000 on 2406 degrees of freedom
## Multiple R-squared:  0.4359, Adjusted R-squared:  0.4024
## F-statistic:    13 on 143 and 2406 DF,  p-value: < 2.2e-16
```

Ridge Regression

The next type of model to try is a ridge regression. For the first model we try an approach by selecting lambda to be really large (equal to 1e10 in this case). In this case the test MSE comes to 6.79e10.

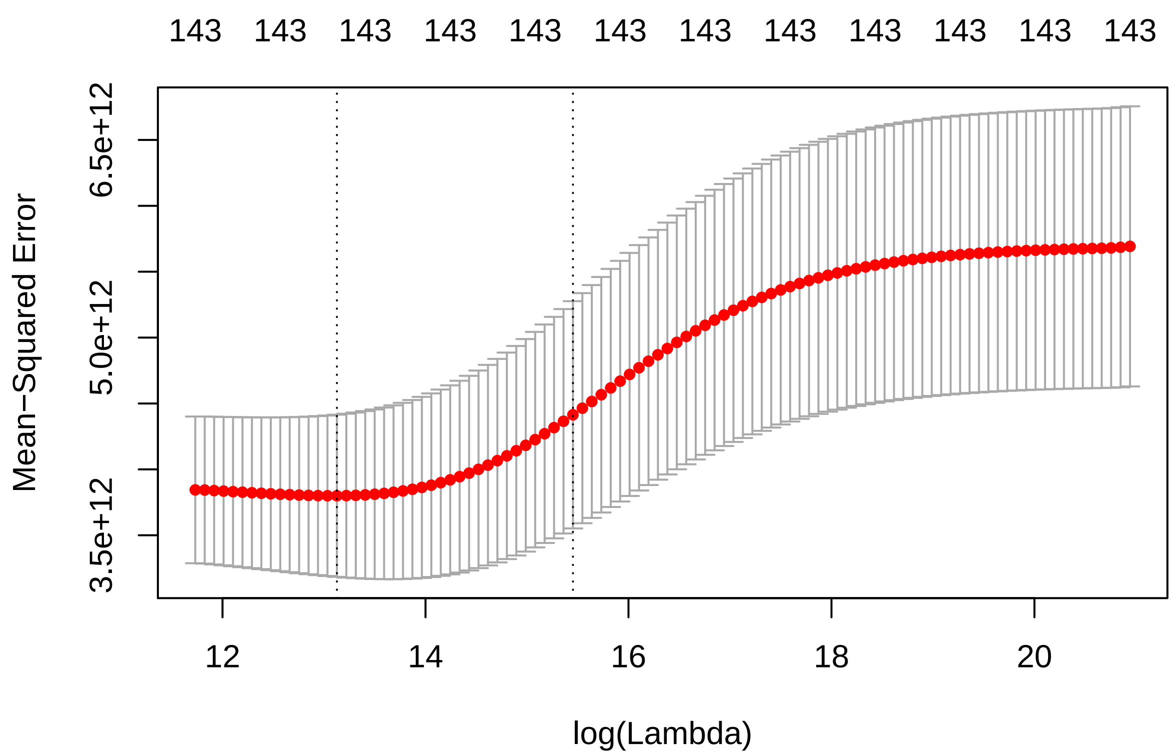
```r
library(glmnet)

set.seed(1)
data <-TedClean
```

```r
y <- as.double(as.matrix(data$views)) # Only class
data$views <- NULL
x <- as.matrix(data) # Removes class

# Fitting the model (Ridge: Alpha = 0)
#select test and train data
set.seed(1)
train <- sample(1:nrow(x),nrow(x)/2)
test <- (-train)
y.test <- y[test]

ridge.mod <-glmnet(x[train,],y[train], alpha = 0)

#ridge regression with really large lambda
ridge.predlarge <- predict(ridge.mod,s=1e10,newx = x[test,])
#test MSE for large lambda
mean((ridge.predlarge-y.test)^2)
```

```
## [1] 6.79394e+12
```

Next, we try ridge regression with cross-validation. The best lambda values is determined to be 502331.5, so we use that value in our model and find that our test MSE for this approach significantly decreases to 4.37e12. This is a high MSE but better than the MSE with the large value of lambda.

```r
#ridge regression with cross validation

#first select best lambda
set.seed(1)
cv.out <- cv.glmnet(x[train,],y[train],alpha = 0)
plot(cv.out)
```



```r
bestlam <- cv.out$lambda.min
bestlam
```

```
## [1] 502331.5
```

```r
#run with best lambda
ridge.pred <- predict(ridge.mod,s=bestlam,newx = x[test,])

#what is test MSE associeated with lambda = bestlam?
mean((ridge.pred -y.test)^2)
```

```
## [1] 4.374891e+12
```

```r
# so the test MSE decreased by half but it is still very large
```

Below, we refit our ridge regression model with the chosen lambda value. As expected none of the coefficients are zero, and the model is highly uninterpretable. Ridge regression was a fun approach to see what would happen, but it doesn't yield any meaningful results. Next we will try a lasso approach, which is expected to help with variable selection.

```r
#Refit ridge regression model on the full data set with cv chosen lambda
set.seed(1)
out <- glmnet(x,y,alpha=0)
predict(out,type = "coefficients",s=bestlam)
```

```
## 146 x 1 sparse Matrix of class "dgCMatrix"
##                                        1
## (Intercept)                  -937646.5604
## comments                        3338.7439
## duration                         425.6581
## languages                      63868.1267
## tags_can                      -32601.6026
## tags_life                     158153.2682
## tags_new                     -143495.2126
## tags_world                    -13788.8231
## tags_future                    11500.0153
## tags_art                     -139112.5282
## tags_make                     305804.4793
## tags_design                  -126640.7308
## tags_brain                    487500.6308
## tags_better                   132143.1561
## tags_love                     340400.7834
## tags_change                  -111166.3562
## tags_story                    157536.6855
## tags_need                    -185681.8350
## tags_science                 -212427.9833
## tags_power                   1010557.7635
## tags_time                    -285385.4577
## tags_human                   -166019.0994
## tags_women                   -200460.0538
## tags_one                      -58918.5995
## tags_technology               -91942.9165
## tags_global                  -275583.7467
## tags_issues                  -306058.2529
## tags_health                   130536.5573
## tags_culture                  -32554.7159
```

```
## tags_tedx                                     .
## tags_business                       318381.9744
## tags_entertainment                  196566.4169
## tags_social                          99775.5044
## tags_ted                            175907.0650
## tags_biology                        -69814.3956
## tags_innovation                     124612.3099
## tags_society                        189196.0343
## tags_music                         -211391.0752
## tags_communication                  253849.8522
## tags_creativity                     193876.0740
## tags_economics                     -311641.6727
## tags_humanity                       196843.7404
## tags_collaboration                  -16158.0684
## tags_environment                   -102402.4462
## tags_medicine                      -122453.0489
## tags_activism                      -159609.6883
## tags_education                      160873.6887
## tags_community                     -113680.5623
## tags_history                        -58314.9759
## tags_children                       -77005.2020
## tags_fellows                                  .
## tags_performance                    329757.0600
## tags_invention                       -8878.9137
## tags_psychology                     916200.3803
## tags_care                          -323349.2076
## tags_politics                      -416726.0188
## tags_cities                        -238700.1970
## tags_energy                        -295994.4355
## tags_media                         -144149.5079
## tags_storytelling                  -163524.8372
## tags_nature                         268625.6904
## tags_war                            -92137.6246
## tags_identity                        11281.1755
## tags_computers                      -84297.2468
## tags_engineering                    -28223.6956
## tags_animals                          3314.6831
## speaker_occupation_Writer           406468.1707
## speaker_occupation_Artist          -223711.3069
## speaker_occupation_Designer        -148718.3754
## speaker_occupation_Journalist      -255370.2653
## speaker_occupation_Entrepreneur    -215584.2765
## speaker_occupation_Architect        -80654.2171
## speaker_occupation_Inventor        -128551.2164
## speaker_occupation_Psychologist     222216.1883
## speaker_occupation_Photographer    -343897.9270
## speaker_occupation_Filmmaker       -158435.7677
## speaker_occupation_Author           324695.4572
## speaker_occupation_Economist        -94234.9836
## speaker_occupation_Educator        -171934.1585
## speaker_occupation_Neuroscientist  -515869.3151
## speaker_occupation_Philosopher     -698567.1339
## speaker_Hans.Rosling               468635.2344
## speaker_Juan.Enriquez                -2998.1630
```

```
## speaker_Marco.Tempest              296994.3811
## speaker_Rives                     -466438.8066
## speaker_Bill.Gates                 -96595.6265
## speaker_Clay.Shirky               -826557.8171
## speaker_Dan.Ariely                 151141.3463
## speaker_Jacqueline.Novogratz      -157102.2513
## speaker_Julian.Treasure           3081202.7769
## speaker_Nicholas.Negroponte       -142582.6196
## speaker_Al.Gore                  -1308406.4459
## speaker_Barry.Schwartz              26390.4189
## speaker_Chris.Anderson            -391891.3978
## speaker_Dan.Dennett               -607636.8254
## speaker_David.Pogue                 13478.2312
## event_TED2014                      122252.7705
## event_TED2009                     -148395.0272
## event_TED2013                      -70198.2618
## event_TED2016                      372494.9637
## event_TED2015                      224438.4999
## event_TED2011                     -299276.5598
## event_TEDGlobal.2012               261582.4525
## event_TED2007                      -13839.0941
## event_TED2010                     -730952.4377
## event_TEDGlobal.2011              -476016.8670
## event_TED2017                      752252.2794
## event_TEDGlobal.2013               344600.4083
## event_TED2012                       -1474.5521
## event_TEDGlobal.2009                36051.9819
## event_TED2008                       64928.1688
## event_TEDGlobal.2010              -971180.1113
## event_TEDGlobal.2014               -69208.8551
## event_TED2006                     1105065.5111
## event_TED2005                      163916.8172
## event_TEDIndia.2009               -477359.4953
## description_talk                   125760.1488
## description_can                     12392.5751
## description_world                   22804.1840
## description_new                   -144562.3884
## description_says                  -138435.3566
## description_shares                 -75337.6703
## description_people                  48670.5070
## description_ted                     30429.2565
## description_shows                  -93957.3090
## description_one                     79447.6027
## description_life                   -65295.9274
## description_like                    99203.5809
## description_make                    65462.8621
## description_way                   -108025.5593
## description_human                 -103914.3482
## description_work                    84271.3897
## description_just                    17006.0215
## description_help                   140681.5644
## description_story                  -59978.3095
## description_even                   118472.9531
## description_time                    86073.2983
```

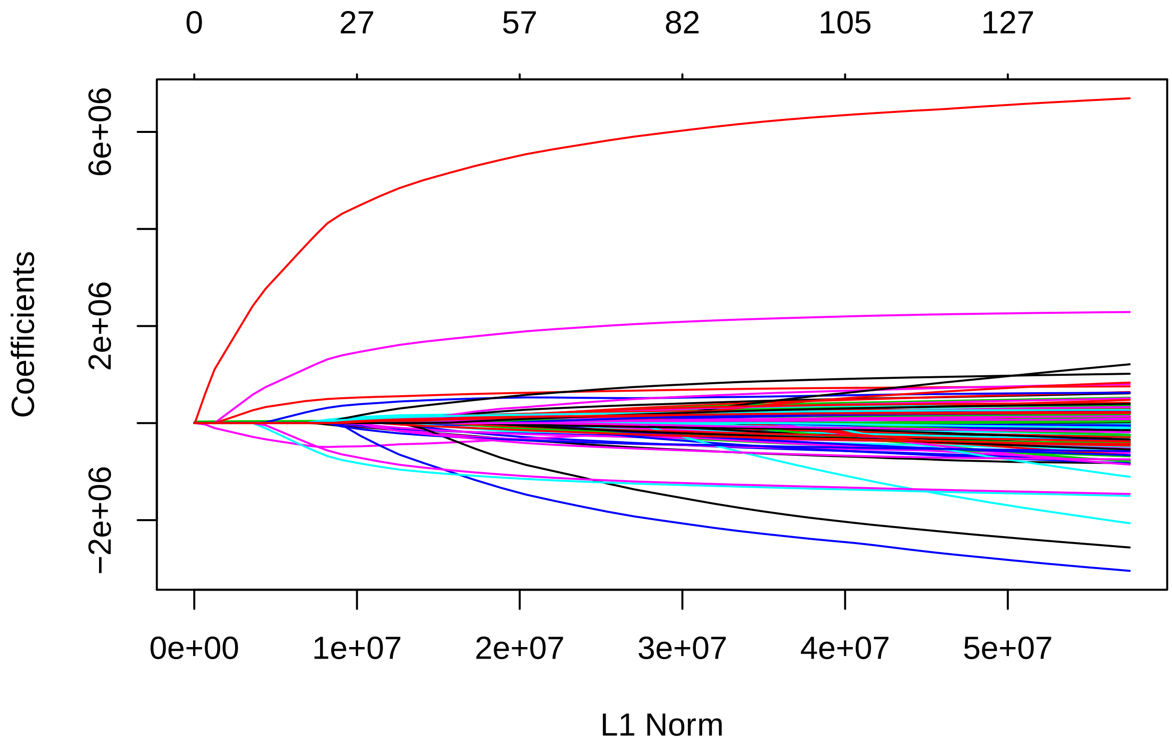
```
## description_years                          4587.4534
## description_makes                       -165695.8518
## description_talks                       -420001.2450
## description_will                         -95185.3917
## description_data                          33751.1236
## description_future                       -61815.4411
## description_change                       149014.9170
## description_powerful                      49987.7948
## description_now                           80845.2788
```

```
##As expect none of the coefficients are zero and this is an highly uninterpretabe model.
```

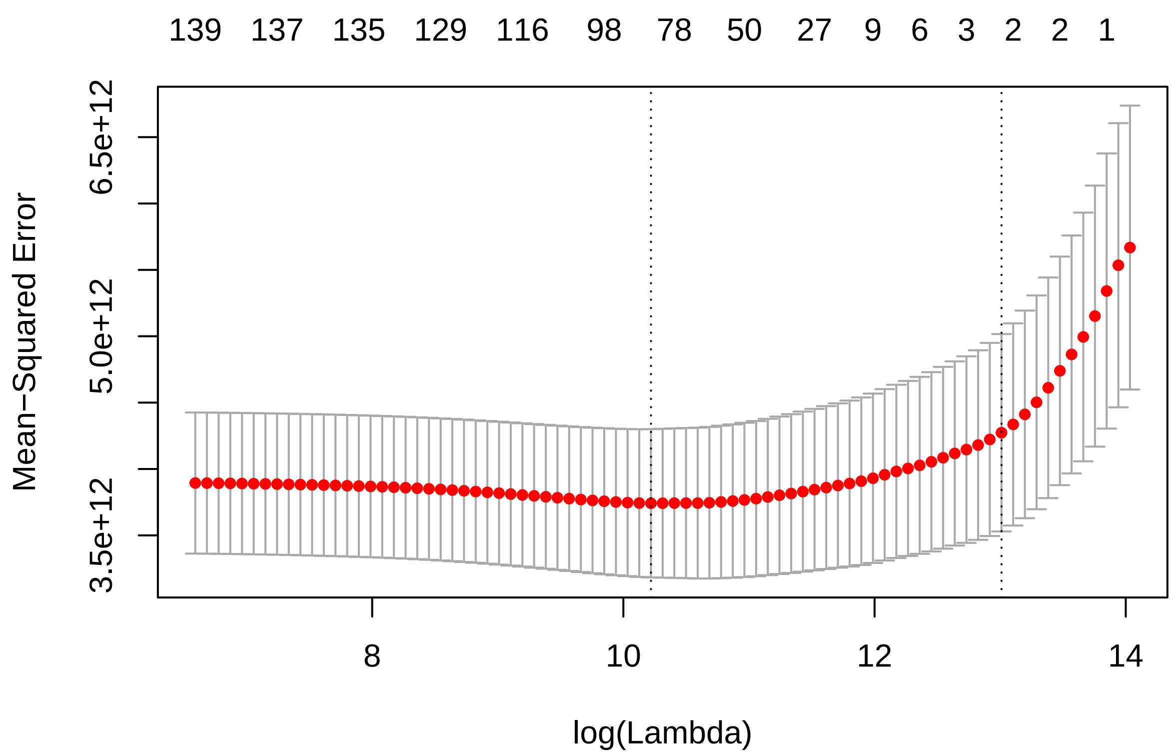
Lasso

Below we use a lasso approach with lambda determined to be 27438.74. We can see the test MSE is 4.24e12 which is lower that the test MSE of ridge regression.

```
library(glmnet)

set.seed(1)
data <- TedClean
y <- as.double(as.matrix(data$views)) # Only class
data$views <- NULL
x <- as.matrix(data) # Removes class


# Fitting the model (Ridge: Alpha = 0)
#select test and train data
set.seed(1)
train <- sample(1:nrow(x),nrow(x)/2)
test <- (-train)
y.test <- y[test]

lasso.mod <- glmnet(x[train,],y[train], alpha = 1)
plot(lasso.mod)
```




```r
#perform cross validation with lasso
set.seed(1)
cv.outLasso <- cv.glmnet(x[train,],y[train],alpha=1)
plot(cv.outLasso)
```




```r
bestlamlasso <- cv.outLasso$lambda.min
bestlamlasso
```

```
## [1] 27438.74
```

```r
lasso.pred <- predict(lasso.mod,bestlamlasso, newx=x[test,])
mean((lasso.pred - y.test)^2)
```

```
## [1] 4.425157e+12
```

```r
#lower test MSE that ridge regression with lambda chosen by cv
```

Below we fit the lasso model. We can see that is has deemed a number of predictors significant while also remove a chunk. The results can be summarized as the following have a positive effect on the number of views: comments, duration, language, tags: life, make, brain, better, love, power, health, business, entertainment, innovation, society, communication, creativity, humanity, performance, psychology, and nature, speaker occupation – writer, author, speaker is Julian Treasure, event is – TED2014, TED2015, TED2016, TEDGlobal.2012, TED2017,description contains – talk, people, ted, one, like, work, help, even, time, change, and now. This model is much more interpretable than the ridge regression model, but does contain a large number of predictors.

```r
#refit lasso
set.seed(1)
outlasso <- glmnet(x,y,alpha=1)
lasso.coef <- predict(outlasso,type = "coefficients", s=bestlamlasso)
lasso.coef
```

```
## 146 x 1 sparse Matrix of class "dgCMatrix"
##                                        1
## (Intercept)                -1344649.3412
## comments                       3932.4741
## duration                        453.4372
## languages                     72248.6035
## tags_can                               .
## tags_life                     57312.1375
## tags_new                               .
## tags_world                             .
## tags_future                            .
## tags_art                     -42607.6359
## tags_make                    209338.2936
## tags_design                  -80597.0594
## tags_brain                   423519.5410
## tags_better                    5340.2358
## tags_love                    241344.6994
## tags_change                            .
## tags_story                             .
## tags_need                    -21595.9170
## tags_science                -212637.0145
## tags_power                   951737.3015
## tags_time                    -73646.8441
## tags_human                             .
## tags_women                  -106493.7088
## tags_one                               .
## tags_technology              -42086.0610
## tags_global                 -172160.0758
## tags_issues                 -446187.8502
## tags_health                     884.5210
## tags_culture                 -49019.1628
## tags_tedx                              .
```

```
## tags_business                           266697.2386
## tags_entertainment                       96234.1553
## tags_social                                   .
## tags_ted                                      .
## tags_biology                                  .
## tags_innovation                          45419.6075
## tags_society                            162966.9536
## tags_music                                    .
## tags_communication                      218408.7220
## tags_creativity                          95568.0144
## tags_economics                         -208156.4698
## tags_humanity                           155240.6196
## tags_collaboration                            .
## tags_environment                              .
## tags_medicine                            -2057.4808
## tags_activism                           -72184.6035
## tags_education                                .
## tags_community                                .
## tags_history                                  .
## tags_children                                 .
## tags_fellows                                  .
## tags_performance                        285587.4773
## tags_invention                                .
## tags_psychology                         964682.9029
## tags_care                              -156153.5968
## tags_politics                          -370325.7675
## tags_cities                            -171005.2012
## tags_energy                            -234501.3043
## tags_media                              -16117.7889
## tags_storytelling                       -41781.5467
## tags_nature                             219061.2659
## tags_war                                      .
## tags_identity                                 .
## tags_computers                                .
## tags_engineering                              .
## tags_animals                                  .
## speaker_occupation_Writer               295437.2811
## speaker_occupation_Artist              -107160.1031
## speaker_occupation_Designer                   .
## speaker_occupation_Journalist           -87508.1799
## speaker_occupation_Entrepreneur         -50865.8161
## speaker_occupation_Architect                  .
## speaker_occupation_Inventor                   .
## speaker_occupation_Psychologist               .
## speaker_occupation_Photographer        -172867.7104
## speaker_occupation_Filmmaker                  .
## speaker_occupation_Author               141174.8330
## speaker_occupation_Economist                  .
## speaker_occupation_Educator                   .
## speaker_occupation_Neuroscientist      -412990.4807
## speaker_occupation_Philosopher         -796060.7794
## speaker_Hans.Rosling                          .
## speaker_Juan.Enriquez                         .
## speaker_Marco.Tempest                         .
```

```
## speaker_Rives                                   .
## speaker_Bill.Gates                              .
## speaker_Clay.Shirky                  -491727.9946
## speaker_Dan.Ariely                              .
## speaker_Jacqueline.Novogratz                    .
## speaker_Julian.Treasure               3012855.6470
## speaker_Nicholas.Negroponte                     .
## speaker_Al.Gore                     -1189410.1839
## speaker_Barry.Schwartz                          .
## speaker_Chris.Anderson                          .
## speaker_Dan.Dennett                   -86368.5078
## speaker_David.Pogue                             .
## event_TED2014                           7718.9890
## event_TED2009                        -151363.1994
## event_TED2013                         -47879.6694
## event_TED2016                         324710.1249
## event_TED2015                         100889.1673
## event_TED2011                        -371644.4049
## event_TEDGlobal.2012                   48983.7408
## event_TED2007                                   .
## event_TED2010                        -921631.9585
## event_TEDGlobal.2011                 -529939.4190
## event_TED2017                         912403.9069
## event_TEDGlobal.2013                  172864.9788
## event_TED2012                                   .
## event_TEDGlobal.2009                            .
## event_TED2008                                   .
## event_TEDGlobal.2010                -1161299.5910
## event_TEDGlobal.2014                            .
## event_TED2006                         997089.5914
## event_TED2005                                   .
## event_TEDIndia.2009                  -409524.3286
## description_talk                       71341.3906
## description_can                                 .
## description_world                               .
## description_new                      -108613.8046
## description_says                      -41151.2733
## description_shares                     -1002.9965
## description_people                      9968.1123
## description_ted                        14984.4857
## description_shows                      -2353.8180
## description_one                        55891.3015
## description_life                                .
## description_like                       71529.9570
## description_make                                .
## description_way                       -24176.0726
## description_human                     -51030.1134
## description_work                       49612.8539
## description_just                                .
## description_help                      120712.7690
## description_story                               .
## description_even                       16319.8547
## description_time                       53678.5701
## description_years                               .
```

```
## description_makes                    -16180.8655
## description_talks                   -290374.3195
## description_will                              .
## description_data                              .
## description_future                            .
## description_change                    29602.7471
## description_powerful                          .
## description_now                       29996.6494
```

Bagging

Now we try a bagged model. Running it on the test set we see that the test MSE is 3.21e12, which is lower than our lasso model. 51.09% of the variables are explain in this model. We can see that by far the 3 most important variables are comments, languages, and duration. Then there are some interesting ones are the event being TED2017,TED2015, or TED2016, the speaker being Julian Treasure and the tag containing brain.

```
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
set.seed(1)
train <- sample(1:nrow(TedClean),nrow(TedClean)/2)
test <- (-train)

Ted.test <-TedClean[-train,"views"]

bag.Ted <- randomForest(views~., data=TedClean, subset = train,
                        mtry = 145, importance =T)
bag.Ted
```

```
##
## Call:
##  randomForest(formula = views ~ ., data = TedClean, mtry = 145,      importance = T, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 145
##
##           Mean of squared residuals: 2.781098e+12
##                     % Var explained: 51.09
```

```
#how does the bagged model (all predictors) perform on the test?
yhat.bag <- predict(bag.Ted,newdata=TedClean[-train,])
mean((yhat.bag -Ted.test)^2)
```

```
## [1] 3.213958e+12
```

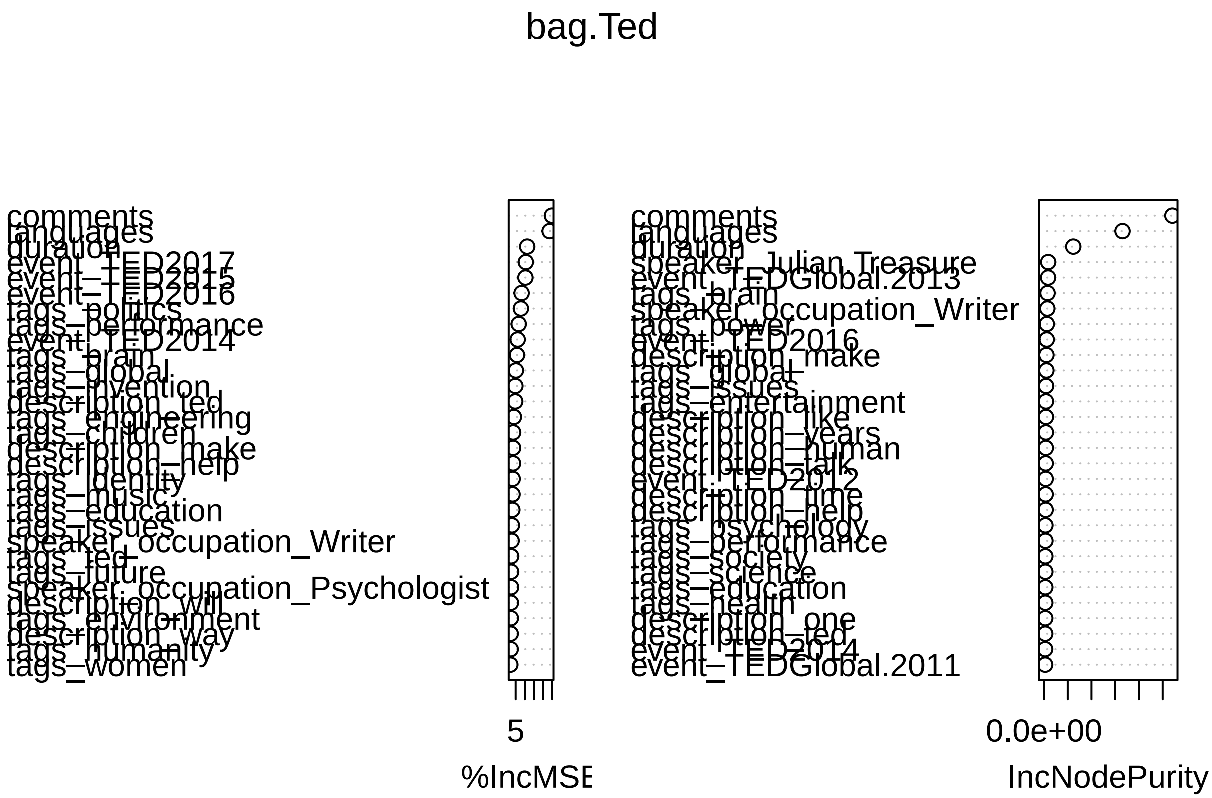
```r
importance(bag.Ted)
```

```
##                                 %IncMSE IncNodePurity
## comments                    24.80293789  2.704863e+15
## duration                    11.30346298  6.170624e+14
## languages                   23.55677670  1.654113e+15
## tags_can                    -1.33098983  4.792185e+12
## tags_life                    0.10458447  4.330830e+12
## tags_new                    -3.04718285  6.248977e+12
## tags_world                  -0.99892796  4.884630e+12
## tags_future                  2.53553497  2.396745e+12
## tags_art                     1.45474682  9.296181e+12
## tags_make                   -1.49156300  5.618963e+12
## tags_design                 -1.64959144  6.546178e+12
## tags_brain                   5.74557879  7.525154e+13
## tags_better                  0.23230289  1.455392e+12
## tags_love                   -6.35906845  2.019227e+13
## tags_change                 -0.49518891  6.356104e+12
## tags_story                  -1.41491268  3.003424e+12
## tags_need                   -2.42558162  1.341207e+12
## tags_science                -0.58988950  2.865127e+13
## tags_power                  -1.25333056  5.950840e+13
## tags_time                    0.30137772  1.977294e+12
## tags_human                   1.36667516  3.291284e+12
## tags_women                   2.09435368  1.389136e+13
## tags_one                     0.39120157  2.539331e+12
## tags_technology             -1.32776524  1.855791e+13
## tags_global                  5.12423527  5.059339e+13
## tags_issues                  2.93777896  4.454832e+13
## tags_health                  1.25034772  2.774586e+13
## tags_culture                -0.94995689  1.752354e+13
## tags_tedx                    0.00000000  0.000000e+00
## tags_business               -0.54142293  1.535082e+13
## tags_entertainment           0.89206229  4.111176e+13
## tags_social                 -1.80912939  2.213592e+13
## tags_ted                     2.54331148  1.391611e+12
## tags_biology                -0.48051496  2.310422e+12
## tags_innovation              1.10182020  4.979809e+12
## tags_society                 0.58852053  2.892314e+13
## tags_music                   3.28881098  1.639161e+13
## tags_communication          -0.50961849  1.494588e+13
## tags_creativity             -1.16716600  1.145176e+13
## tags_economics               0.20124097  4.142615e+12
## tags_humanity                2.29063270  8.007114e+12
## tags_collaboration           0.53023422  4.150199e+12
## tags_environment             2.36873763  5.570167e+12
## tags_medicine                0.28903931  1.524218e+12
## tags_activism               -1.10771277  8.927848e+11
## tags_education               3.18169744  2.834229e+13
## tags_community               2.04832708  3.854608e+12
## tags_history                -0.76761838  2.257629e+12
## tags_children                3.61907018  2.416179e+13
## tags_fellows                 0.00000000  0.000000e+00
```

```
## tags_performance                    6.65896575  3.029358e+13
## tags_invention                      4.85055274  5.648325e+12
## tags_psychology                     0.16179748  3.042045e+13
## tags_care                          -1.77948265  6.243273e+11
## tags_politics                       7.85483784  2.737101e+12
## tags_cities                        -2.72280525  6.974851e+11
## tags_energy                        -1.65248839  3.667488e+12
## tags_media                         -2.77831845  6.863523e+11
## tags_storytelling                  -1.19692207  8.999907e+12
## tags_nature                         0.49379754  1.316923e+13
## tags_war                           -1.99659515  5.350891e+12
## tags_identity                       3.35688676  1.045741e+13
## tags_computers                     -0.79165592  1.623857e+12
## tags_engineering                    4.03928485  1.841821e+12
## tags_animals                       -1.33483722  1.045191e+13
## speaker_occupation_Writer           2.78845206  7.050653e+13
## speaker_occupation_Artist          -0.08166414  6.243287e+11
## speaker_occupation_Designer         0.82015530  1.000229e+12
## speaker_occupation_Journalist      -4.06397059  1.490551e+12
## speaker_occupation_Entrepreneur    -1.68894259  3.612291e+12
## speaker_occupation_Architect        0.40467425  2.001131e+12
## speaker_occupation_Inventor        -0.42930032  1.735547e+12
## speaker_occupation_Psychologist     2.50189818  2.347483e+13
## speaker_occupation_Photographer    -3.81504279  2.001154e+12
## speaker_occupation_Filmmaker       -0.96769461  8.166340e+11
## speaker_occupation_Author          -0.59687417  7.896350e+12
## speaker_occupation_Economist        0.43662244  6.045418e+10
## speaker_occupation_Educator        -0.78454154  8.452304e+11
## speaker_occupation_Neuroscientist -2.66587928  3.900706e+12
## speaker_occupation_Philosopher     -1.13454064  1.690212e+13
## speaker_Hans.Rosling              -1.05799846  1.934545e+12
## speaker_Juan.Enriquez              0.69076504  2.853356e+12
## speaker_Marco.Tempest              2.00083375  8.711086e+11
## speaker_Rives                     -3.80763304  4.552981e+11
## speaker_Bill.Gates                 0.00000000  9.974099e+09
## speaker_Clay.Shirky                0.00000000  1.446863e+09
## speaker_Dan.Ariely                -1.38966201  5.685449e+11
## speaker_Jacqueline.Novogratz      -1.62319320  2.195190e+10
## speaker_Julian.Treasure           -2.42469314  8.863692e+13
## speaker_Nicholas.Negroponte       -1.30826296  6.248279e+10
## speaker_Al.Gore                   -1.00100150  3.073038e+12
## speaker_Barry.Schwartz             0.00000000  1.108511e+10
## speaker_Chris.Anderson             0.00000000  6.309096e+08
## speaker_Dan.Dennett               -1.00100150  4.621194e+10
## speaker_David.Pogue                0.00000000  7.174915e+09
## event_TED2014                      6.10087730  2.461102e+13
## event_TED2009                     -0.26948231  7.265200e+12
## event_TED2013                     -1.96992032  1.278975e+13
## event_TED2016                      8.24690114  5.823439e+13
## event_TED2015                     10.33858777  4.610760e+12
## event_TED2011                     -2.42692075  3.088209e+12
## event_TEDGlobal.2012              -1.47303963  1.652506e+12
## event_TED2007                     -0.70026697  9.919503e+12
## event_TED2010                      0.28869738  7.337160e+12
```

```
## event_TEDGlobal.2011              -1.23135088  2.455597e+13
## event_TED2017                     10.53997041  6.452348e+12
## event_TEDGlobal.2013               0.97548373  8.645108e+13
## event_TED2012                     -3.06912281  3.551670e+13
## event_TEDGlobal.2009              -0.36602667  5.964378e+12
## event_TED2008                     -1.06891318  1.360282e+13
## event_TEDGlobal.2010               1.58319769  2.175695e+12
## event_TEDGlobal.2014               1.45482886  1.159629e+12
## event_TED2006                     -3.74545748  1.276323e+13
## event_TED2005                      0.38379625  2.092832e+12
## event_TEDIndia.2009                1.80773245  4.654167e+12
## description_talk                   0.19319993  3.665186e+13
## description_can                   -0.15686785  1.101140e+13
## description_world                 -2.45907428  1.475684e+13
## description_new                    1.22624646  1.073065e+13
## description_says                   2.07991552  8.068879e+12
## description_shares                -1.74540092  6.962930e+12
## description_people                -0.96203752  6.040644e+12
## description_ted                    4.76085653  2.655074e+13
## description_shows                 -1.79329227  1.370988e+13
## description_one                   -1.63388717  2.708624e+13
## description_life                  -0.12680097  1.576034e+13
## description_like                  -0.05294580  4.097720e+13
## description_make                   3.58991922  5.160081e+13
## description_way                    2.29948259  9.416204e+12
## description_human                 -2.11208015  3.803866e+13
## description_work                   1.68913126  1.516987e+13
## description_just                   0.40616907  9.374567e+12
## description_help                   3.57356416  3.514653e+13
## description_story                  1.00122103  1.520086e+13
## description_even                  -3.82439657  1.701962e+13
## description_time                   1.72454592  3.551520e+13
## description_years                  1.13646238  4.002321e+13
## description_makes                 -0.79295900  3.228404e+12
## description_talks                 -0.55908687  3.929070e+12
## description_will                   2.42156473  2.199826e+13
## description_data                  -1.07037007  2.392440e+13
## description_future                 0.98417518  2.841333e+12
## description_change                -1.58435744  1.414556e+12
## description_powerful              -1.85018814  1.942023e+13
## description_now                   -0.98011055  1.273338e+13
```

```r
varImpPlot(bag.Ted)
```



Random Forest

Now we try the random forest approach with p/3 variables. We can see the test MSE is 3.33e12 still lower than ridge regression and lasso but not lower than the bagged model. Again, we see that comments, languages, and duration are the most important variables. A few others that stand out are: event being TED2015 or TED2017, the speaker being Julian Treasure, the tags: power, performance, global, issues, brain, invention, and society.

```
# try random forest with p/3 variables
set.seed(1)
rf.Ted <- randomForest(views~., data=TedClean, subset = train,
                       importance = T)

rf.Ted
```

```
##
## Call:
##  randomForest(formula = views ~ ., data = TedClean, importance = T,      subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 48
##
##           Mean of squared residuals: 2.988217e+12
##                     % Var explained: 47.45
```

```
yhat.rf <- predict(rf.Ted,newdata=TedClean[-train,])
#much lower MSE error, random forest showed an improvement over bagging
mean((yhat.rf-Ted.test)^2)
```

```
## [1] 3.338326e+12
```

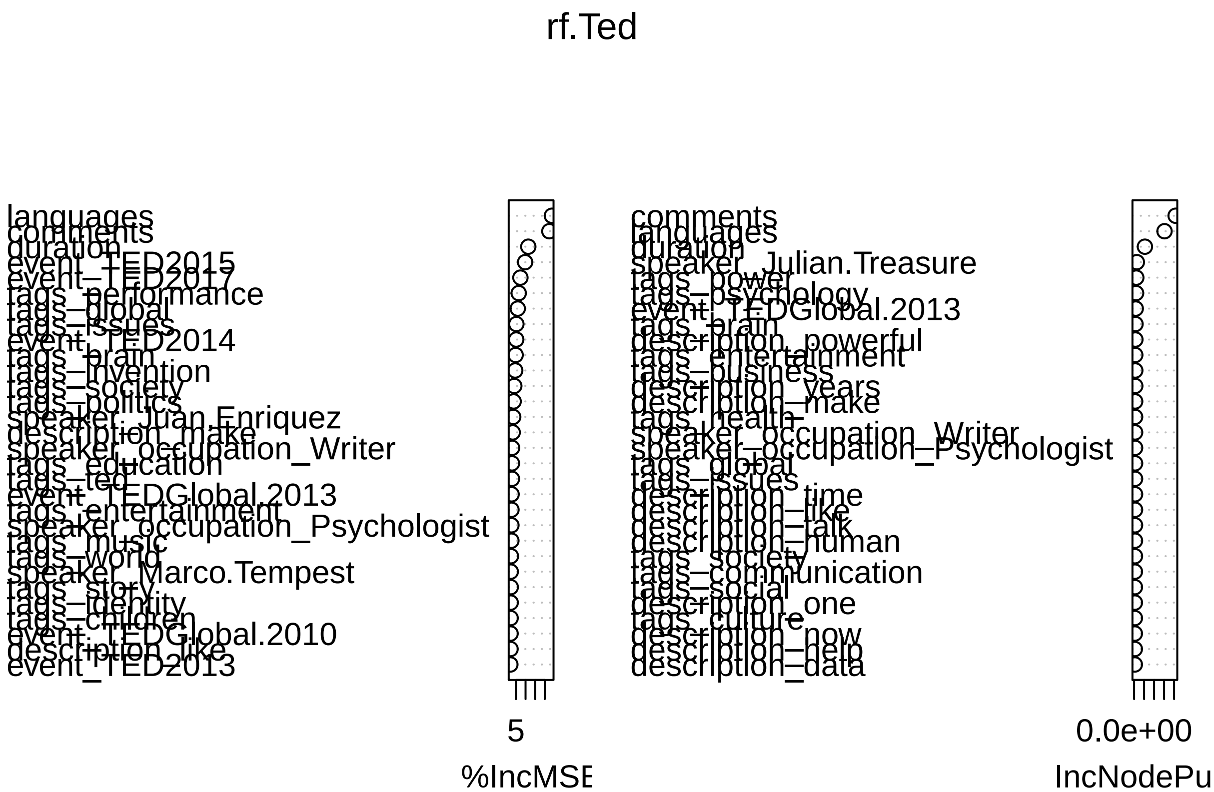
```r
importance(rf.Ted)
```

```
##                                %IncMSE IncNodePurity
## comments                  22.386000030  2.076396e+15
## duration                  11.425342085  5.395609e+14
## languages                 23.701679529  1.520328e+15
## tags_can                  -0.618285916  4.833501e+12
## tags_life                 -1.003018794  1.499220e+13
## tags_new                  -1.266013897  7.363960e+12
## tags_world                 2.380591472  4.670073e+12
## tags_future                0.263050209  3.086871e+12
## tags_art                   0.713345919  1.351240e+13
## tags_make                  1.873843117  1.395456e+13
## tags_design                0.690140126  1.281324e+13
## tags_brain                 4.838736541  7.056155e+13
## tags_better                0.768038798  1.155777e+13
## tags_love                 -0.093321684  1.863186e+13
## tags_change               -0.293129244  2.969815e+13
## tags_story                 2.286172470  5.410856e+12
## tags_need                 -0.082230011  2.858500e+12
## tags_science              -0.347059791  2.714402e+13
## tags_power                -3.177710917  1.094907e+14
## tags_time                 -1.819319336  3.055006e+12
## tags_human                 0.539241089  3.902825e+12
## tags_women                 0.611213081  1.095132e+13
## tags_one                  -1.942029154  3.950762e+12
## tags_technology           -0.020222506  1.930895e+13
## tags_global                5.945567434  4.929655e+13
## tags_issues                5.229723270  4.792358e+13
## tags_health                0.250563781  5.071258e+13
## tags_culture              -1.386530908  3.816849e+13
## tags_tedx                  0.000000000  0.000000e+00
## tags_business             -0.086725586  5.856346e+13
## tags_entertainment         2.671274631  5.906329e+13
## tags_social                1.356749682  3.946394e+13
## tags_ted                   2.922394705  1.155426e+12
## tags_biology               0.969411492  4.645594e+12
## tags_innovation            1.739062108  1.038849e+13
## tags_society               4.140115477  4.228504e+13
## tags_music                 2.633800816  1.793604e+13
## tags_communication         1.072986325  4.050973e+13
## tags_creativity           -0.533441839  1.557131e+13
## tags_economics             1.850440135  4.229721e+12
## tags_humanity              2.084134394  1.469908e+13
## tags_collaboration        -0.314282489  5.407948e+12
## tags_environment           1.318074910  7.181922e+12
## tags_medicine              1.733088002  2.980451e+12
## tags_activism              0.234317084  1.482477e+12
## tags_education             2.925496039  1.823748e+13
## tags_community             1.372603422  4.727335e+12
## tags_history              -0.281804781  2.421514e+12
## tags_children              2.227724085  1.400084e+13
## tags_fellows               0.000000000  0.000000e+00
```

```
## tags_performance                    6.436694507  2.555460e+13
## tags_invention                      4.619754975  4.465401e+12
## tags_psychology                    -0.026105508  9.959020e+13
## tags_care                          -0.494742534  9.340725e+11
## tags_politics                       3.790787425  4.999966e+12
## tags_cities                        -1.458490067  1.353129e+12
## tags_energy                        -1.032764900  5.204463e+12
## tags_media                         -1.499500751  1.386826e+12
## tags_storytelling                  -2.346288240  1.226038e+13
## tags_nature                         1.390191708  1.014110e+13
## tags_war                           -0.132135946  6.926928e+12
## tags_identity                       2.270185528  1.223538e+13
## tags_computers                     -2.910759474  2.498073e+12
## tags_engineering                    1.108784685  1.859245e+12
## tags_animals                        1.080878047  1.531463e+13
## speaker_occupation_Writer           2.995251463  5.029968e+13
## speaker_occupation_Artist          -0.928608584  7.548567e+11
## speaker_occupation_Designer        -0.598245011  1.844233e+12
## speaker_occupation_Journalist      -2.325359179  2.952700e+12
## speaker_occupation_Entrepreneur     2.075949156  5.549642e+12
## speaker_occupation_Architect       -0.327068871  1.907407e+12
## speaker_occupation_Inventor        -1.416489488  1.987799e+12
## speaker_occupation_Psychologist     2.640697939  4.947055e+13
## speaker_occupation_Photographer    -2.878204664  1.800901e+12
## speaker_occupation_Filmmaker       -1.549748360  1.185073e+12
## speaker_occupation_Author          -0.327634669  8.220231e+12
## speaker_occupation_Economist        1.985869746  1.377496e+11
## speaker_occupation_Educator         0.637539681  8.258048e+11
## speaker_occupation_Neuroscientist   1.336260622  3.332747e+12
## speaker_occupation_Philosopher      0.378526646  1.367302e+13
## speaker_Hans.Rosling                0.070694865  6.748926e+12
## speaker_Juan.Enriquez               3.575115380  2.344967e+12
## speaker_Marco.Tempest               2.302089883  1.217532e+12
## speaker_Rives                      -2.547744073  5.785600e+11
## speaker_Bill.Gates                  0.000000000  2.905739e+11
## speaker_Clay.Shirky                 0.000000000  4.977549e+09
## speaker_Dan.Ariely                  0.644209050  1.102844e+12
## speaker_Jacqueline.Novogratz       -0.954345038  2.216674e+10
## speaker_Julian.Treasure            -2.699123496  1.376049e+14
## speaker_Nicholas.Negroponte        -0.455923463  1.016858e+11
## speaker_Al.Gore                     1.365919547  3.756968e+12
## speaker_Barry.Schwartz              0.000000000  5.631664e+10
## speaker_Chris.Anderson              0.000000000  2.281930e+10
## speaker_Dan.Dennett                 0.777400548  6.858598e+11
## speaker_David.Pogue                 0.000000000  8.119436e+10
## event_TED2014                       5.135087396  1.721217e+13
## event_TED2009                       0.039873392  1.099077e+13
## event_TED2013                       2.087833181  1.572108e+13
## event_TED2016                       0.937711465  2.868806e+13
## event_TED2015                       9.779573207  4.142015e+12
## event_TED2011                      -0.405948404  7.127374e+12
## event_TEDGlobal.2012               -1.908280615  2.707314e+12
## event_TED2007                      -1.869658128  1.654647e+13
## event_TED2010                       1.798114012  4.741189e+12
```

```
## event_TEDGlobal.2011            0.096059047  3.345380e+13
## event_TED2017                   7.314698903  3.600207e+12
## event_TEDGlobal.2013            2.729806774  8.192793e+13
## event_TED2012                  -0.363346407  2.752801e+13
## event_TEDGlobal.2009            1.010798987  2.904971e+13
## event_TED2008                   0.727343588  1.205222e+13
## event_TEDGlobal.2010            2.199743247  6.345706e+12
## event_TEDGlobal.2014           -0.291037607  1.439334e+12
## event_TED2006                  -1.746951439  1.540752e+13
## event_TED2005                  -1.040102012  2.825815e+12
## event_TEDIndia.2009            -0.008613933  1.764998e+13
## description_talk                1.323727394  4.646017e+13
## description_can                -1.498971622  1.286561e+13
## description_world              -1.617713106  1.588755e+13
## description_new                -0.690317704  1.223830e+13
## description_says               -0.140265981  1.214057e+13
## description_shares             -0.633447518  2.549151e+13
## description_people              0.066673202  9.984908e+12
## description_ted                -0.148325320  2.714928e+13
## description_shows               0.971753373  3.320105e+13
## description_one                 0.741792969  3.924624e+13
## description_life               -1.037333451  1.678806e+13
## description_like                2.185291853  4.716423e+13
## description_make                3.315196993  5.360310e+13
## description_way                 0.176850132  1.412287e+13
## description_human              -1.574341107  4.349002e+13
## description_work               -1.175795681  1.922578e+13
## description_just                0.929214023  1.164123e+13
## description_help                0.816147646  3.649458e+13
## description_story               1.469006620  1.739749e+13
## description_even               -0.499749824  1.985367e+13
## description_time                1.683002947  4.776229e+13
## description_years               0.887449913  5.855543e+13
## description_makes              -0.522902952  5.177863e+12
## description_talks               1.828435373  5.582439e+12
## description_will                0.235312396  1.512388e+13
## description_data                1.386128160  3.552855e+13
## description_future              0.729382021  5.656018e+12
## description_change              0.256372528  3.425094e+12
## description_powerful           -0.438879058  6.454658e+13
## description_now                -1.287690650  3.791678e+13
```

```
varImpPlot(rf.Ted)
```



Conclusion

Five approaches were tried to determine what predictors had significant effects on the number of views a TED Talk received. The models all had pretty high test MSE, and while a bagged approach fit the data best out of all the models it was not an objectively great fit. However, there were a few predictors that stood out amongst multiple models which leads me to conclude that they are significant. The ridge regression model was largely uninterpretable so not much was conclude much from it. The predictors comments, duration, and languages were all top predictors in in the Linear, Lasso, Bagged, and Random Forest therefore we can be fairly confident those are significant. In both the Linear and Lasso model the tags "art" and "design" were significant in decreasing the number of views. In the Lasso and Bagged models the tag "brain" was shown to increase the number of views. Finally, the speaker Julian Treasure was deemed to significantly increase the number of views in the Lasso and Bagged models. The findings can be summarized by saying an increase in the comment number, length of the talk, and languages translated into are good indicators of an increase in the number of views the talk receives. As far as the content, talks tagged "brain" receive more views, and talks tagged "art" and "design" tend to receive less views. The speaker Julian Treasure is shown to command a significant number of views.

In terms of improving the models, having more data would go a really long way. While, there are 2550 talks in the dataset, the data was split in half for training and testing. It was a difficult choice between training on more data or doing an even split, but the dataset contains a lot of really unique talks and ultimately a lot of predictors. It is entirely plausible that whether one really popular talk is in the training or test dataset actually makes a difference. The original 17 columns were transformed into 145 predictors and could have been made into more. Additionally, pulling in more data could have helped with any amount of overfitting that was happening.

While this problem was approached as a regression problem, in the future it could be interesting to try classification and see if that gleaned any interesting results. It might lead to something that would be difficult to interpret, but could show more precise classes.