STAD37: Multivariate Analysis

Final Project

Group Number: 9

Teammate Last Name, First name and Student Number:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Last Name | First Name | Student Number |
| 1. | EDWARDS | JOHN |  |
| 2. | KAN | KAMILLA |  |
| 3. | SON | JONGHWAN | 999959843 |

Table of Contents

**Description of Research Question1**

**Description of Scientific or Research Issues1**

**Description of Data1**

**Statistical Analysis2**

Logistic Regression Analysis**8**

STAD37 Final Project

# Description of Research Question:

The primary goal of this analysis is to assess the popularity of TED Talks and identify the associated characteristics that predict this measure. Furthermore, we are interested in comparing how predictors change with respect to a TED Talks’ topic and time of publication.

# Description of Scientific or Research Issues:

The data initially contained a large number of variables, which made it difficult to identify the significant variables. Furthermore, after some preliminary analysis, we discovered that the nature of the data is not normally distributed, so it limited our ability to implement statistical inference. Additionally, the nature in which the variables were recorded is ambiguous which poses the risk of bias in their measurements.

Certain variables provided limited predicting power because they are only recorded once the talk has been published. It was our view that in order to predict the popularity of a TED Talk, we would want to consider its state of characteristics before release. Furthermore, the subjective definition of popularity poses ambiguity in a definitive measure. As investigators, we have our own opinions on what constitutes popularity. However, different audiences may vary in their opinion of this.

# 

# Description of Data:

In order to assess how various characteristics of TED talks change with time, we needed to convert the provided publish dates into a numerically readable format that would allow observations to be classified by day, month and year. We opted not to assess the film date as we believed the publication date more accurately reflected the audience’s opinions at that time. Below we created 3 variables of day, month, and year to assess trends over time throughout our analysis.

A characteristic we hypothesized could predict popularity was the view count of related videos. Related talks appear at the bottom of the web page; if the related talks are popular (measured by the view count), there is a high chance that after watching them people would also watch related talks. We decided to take a count of views for related talks for each individual Ted Talk and take the average of that value.

**Create Sentiment Scores and Counts:** In order to easily evaluate how ratings influence popularity, we decided to create the sentiment score system. This was done by observing the number of times a talk has received a certain evaluation (i.e. “Persuasive”, “Informative”, etc.…), determining whether it is negative or positive, and counting evaluations per number of views. Furthermore, we recorded the individual counts of these ratings to explore if a specific feeling toward a talk better predicted its popularity.

The “Tags”, “Title”, and “Description” variables contain a large amount of text. We believed that the subject nature of a video could be determined by assessing a combination of these columns. In order to employ a variable which could be used to make inferences about popularity, we attempted to group videos into topic groups based on a combination of this text data.

This process was accomplished through a Latent Dirichlet Analysis Topic Model. In order to conduct this analysis, we needed to prepare the data for the LDA model. This was accomplished by removing all the unnecessary tags, such as “TED Fellows”, “TEDx”, “TED Brain Trust”, “TED MEd” and “talks” since they do not have a significant role in grouping topics. Additionally, we removed stop words such as “and”," “if”, “or”, etc.… which did not provide interactability toward a specific topic. Finally, we removed words with less than 5 occurrences amongst all video text. We then created the document term matrix with the frequency of each unique word for each video. This was then implemented into the LDA Topic Model which will be discussed later in the analysis.

While removing unuseful text we were able to create a new variable of event type between TED Talks and TEDx. We opted to use this variable rather than the event variable in the original dataset because it is not as diluted by fragmented categories like the city or event locations in the original event variable. Because TED and TEDx are intended for different kind of public (TED covers more global issues and focuses on the global community, while TEDx is typically intended for a local community and addresses local issues), the distribution of popular videos among these two groups might differ as well.

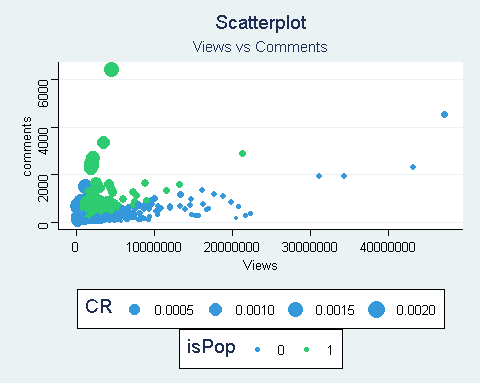
In predicting the popularity of TED talks there are a variety of ways in which popularity can be defined. Total views act as an estimate of the size of the audience the video has been able to reach. However, there is no detailed information on how long viewers watched the video for or whether they left soon after starting a video. To combat this ambiguity, we opted to include a measure of user activity in our classification of popularity

We defined a new measure called the comment ratio which is the total number of comments divided by the total views of a video. The intuition behind this statistic was to judge user activity relative to the number of views. It was then decided that a talk could be deemed popular if it had both above median views and an above median comment ratio. Our logic here was, if it measured high on both these stats, it meant many individuals were viewing the video and engaging in discussion about it.

We considered including the sentiment score calculation into this measure. However, we found that it did not offer clear differentiation between videos. we wanted the opportunity to assess the effect of individual rating types on popularity without introducing a bias towards it in the measure.

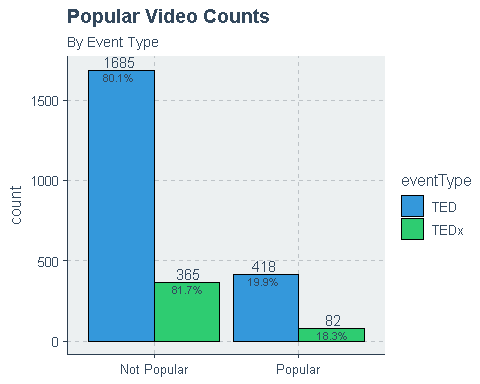
Implementing this classification rule led to the following segmentation of the data. As we can see from the scatterplot, there is a high correlation between views and comments. Thus, we took the comment ratio as one of our main variables to classify the video as popular or not popular. Here, the variable isPop was defined.

# Statistical Analysis

This revealed the following structure of Popular Vs. Not Popular counts. We noticed that combining the elements of view counts and relative user activity by comment ratios distinguished a more elite class of 500 videos whose characteristics we now wish to explore.

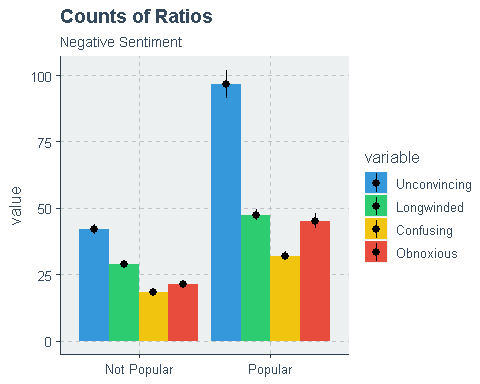
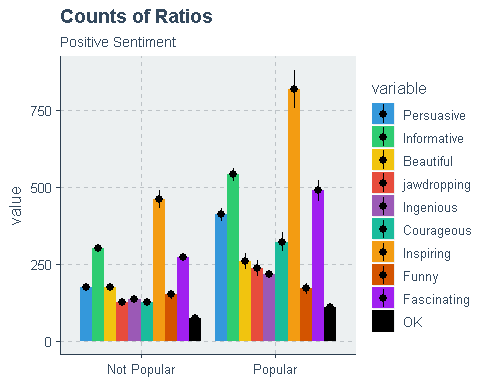
Our next goal was to determine the features that led to these classifications. We first attempted a factor analysis of the quantitative variables to explore the possibility of variable reduction. However, as seen below this was unsuccessful in revealing a simple structure to the data with two factors explaining a small portion of the variance in both Bartlett’s and the Regression methods calculation of factors. Furthermore, when testing the null hypothesis that a sufficient number of factors was used in our factor analysis, the test yielded p-values close to zero.  
## Factor1 Factor2  
## SS loadings 141.750 141.420  
## Proportion Var 5.906 5.892  
## Cumulative Var 5.906 11.799

Our next approach was to assess how popularity differed with different categorical variables. We first assessed how the counts of popular videos differed by eventTypes previously defined above (TED, TEDx). We did not see a sizable difference between the proportions of popular videos amongst the categories. Perhaps further segmentation later in the analysis will reveal a more interesting structure amongst these types.

 Here, we compared the means of views, comment ratios, counts of Ratings, and counts of popular talks between the event types using MANOVA. Using Wilk’s lambda, we reject the null hypothesis that the covariance matrices are equal. Thus, inference between means is limited.

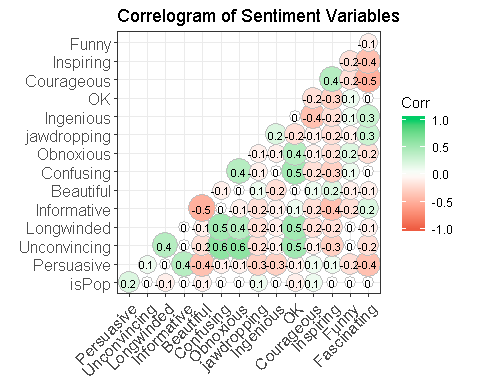
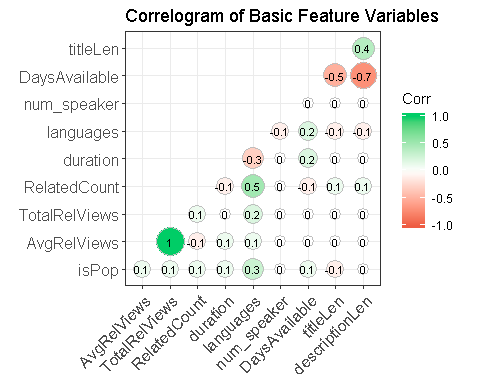
## Df Wilks approx F num Df den Df Pr(>F)   
## eventType 1 0.99029 6.2367 4 2545 0.0000543 \*\*\*  
## Residuals 2548

We further visualized the difference of ratings values between popularity by plotting the respective counts of rating types between Popular and Not Popular segments. We first highlight that the popular videos have larger counts in all rating types. This provides evidence that having larger views and comment activity corresponds with a higher number of ratings. Amongst the positive rating types, the “Inspiring” rating has a significantly greater count in both popularity groups. Amongst the negative counts, the “Unconvincing” rating shares this property of the highest count amongst popularity groups. This provides evidence of a crowding effect of perceptions by viewers. If there is already a majority of people tending towards a rating type, other viewers may also tend toward this sentiment.



In order to better understand the relative effects each rating variable count has on its respective video, we decided to convert rating counts to their proportions amongst the total number of ratings for each respective video.

To get a better sense of the relationship between popularity and the feature variables, we assessed the correlation matrix for the quantitative variables.



This raises some interesting findings. Languages having the highest correlation with popularity and could indicate allowing the video to be understood by a more global audience is resulting in larger groups viewing and interacting on videos. There is also a variety of minor positive and negative correlations amongst the ratings ratios. Before jumping to any conclusions, it important that we consider how these effects change based on segments of categorical variables in addition to time.

On the graph Average number of views that are distributed all over years 2007 to 2018 we can see that the highest number of views occurred in 2007, following by a huge decline and a small spike in 2013, again following the downwards trend to 2018. This can be backed up by the fact that Ted platform was launched in 2007( that is why we can observe such a huge number of views), and in March 2012 Netflix announced a deal to stream series of Ted Talks(that explains a small upward trend in the number of views, since expansion of Ted Talks to more platforms provides access for bigger amount of population).

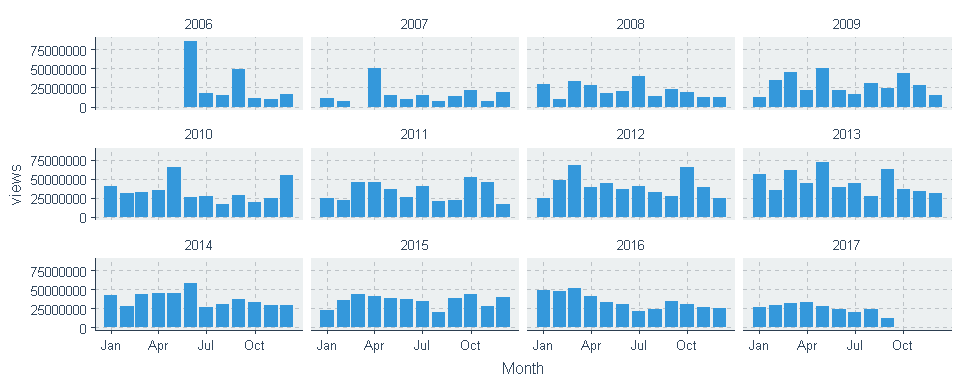
The graph average comments distribution per year indicates that there were major spikes of an upward trend in 2011 and in 2014. Although, the nature of the trends is not clear, we can theorize that the spike in comments could be influenced by some amount of talks that were appealing for majority of people during that time (could be caused by political or media influence). This theory could be backed up by the next graph that indicates that the most popular videos also appeared on 2011 and 2014, which supports our choice of including the comment per view variable into our model that defines the popularity.

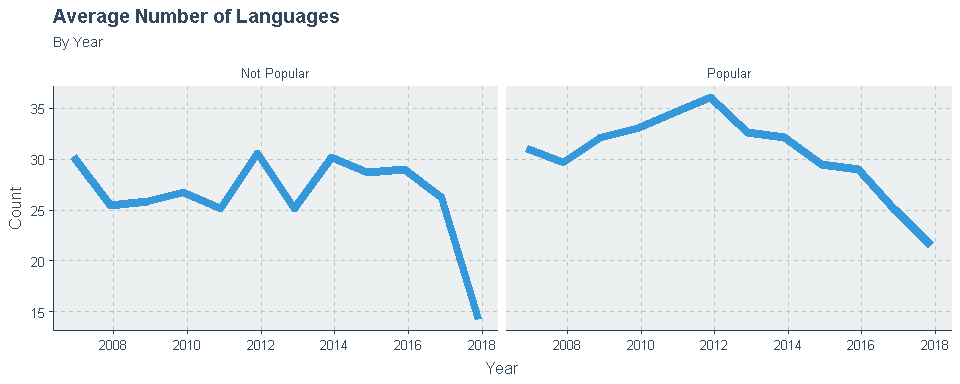
Average number of languages by years for both popular and unpopular videos show an overall decreasing trend. This is believed to be because more recent videos have not had sufficient time to be translated into multiple languages.

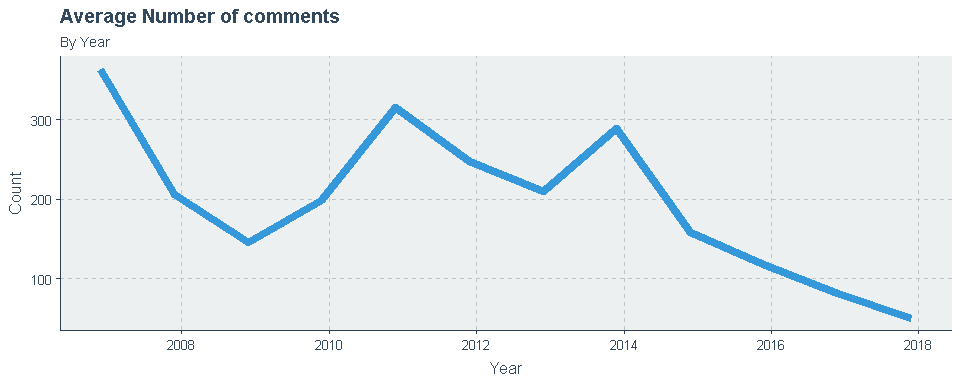
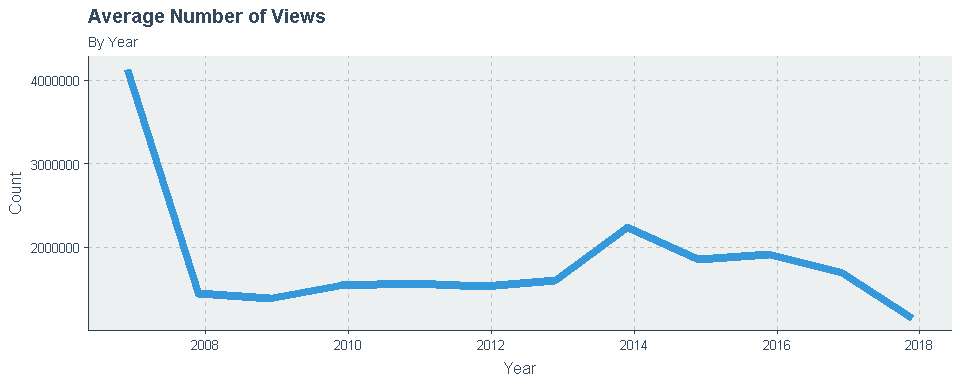
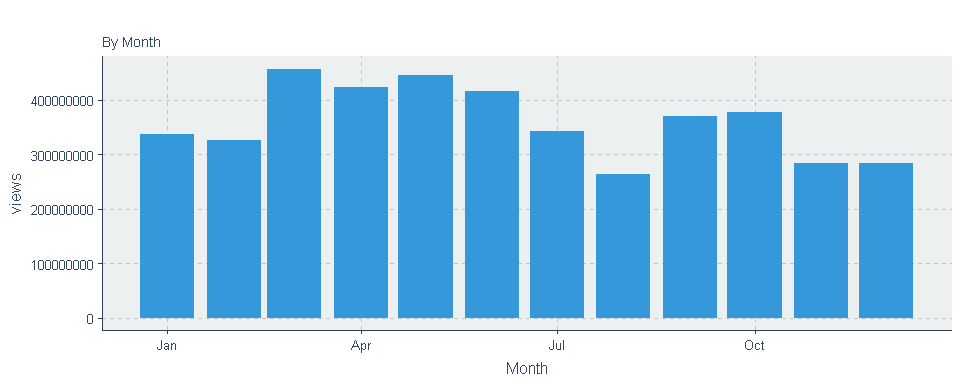
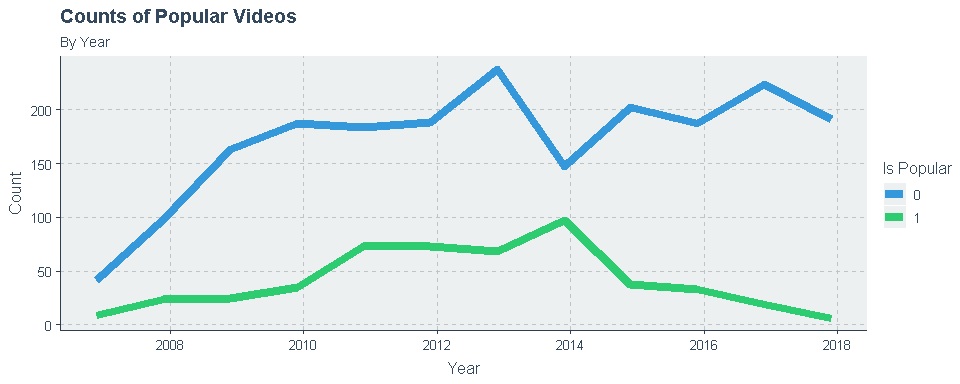
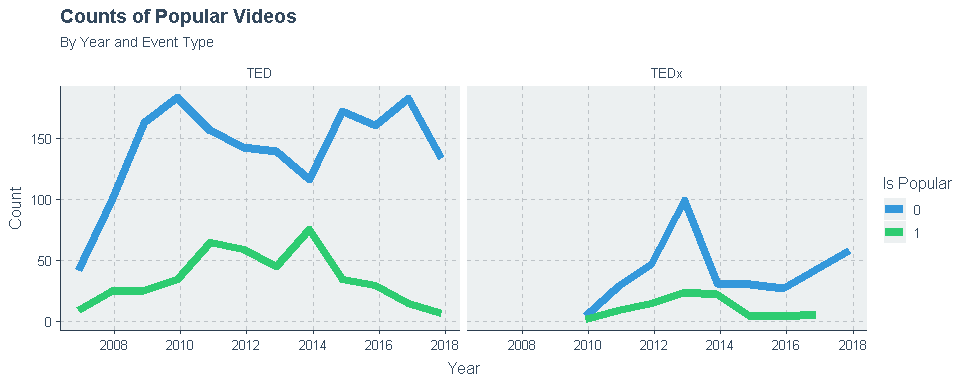
Number of views by month shows that there are less views, averaging around 300000000, in the winter months. There is an increase in views in March and a gradual decline into the summer months, ending in August. It is difficult to infer a particular trend in views by month because the audience varies so greatly. Furthermore, the count of popular and unpopular videos published by month does not vary greatly. Therefore, it cannot be inferred that number of views in any given month increases or decreases with the number of videos that were published in that month.

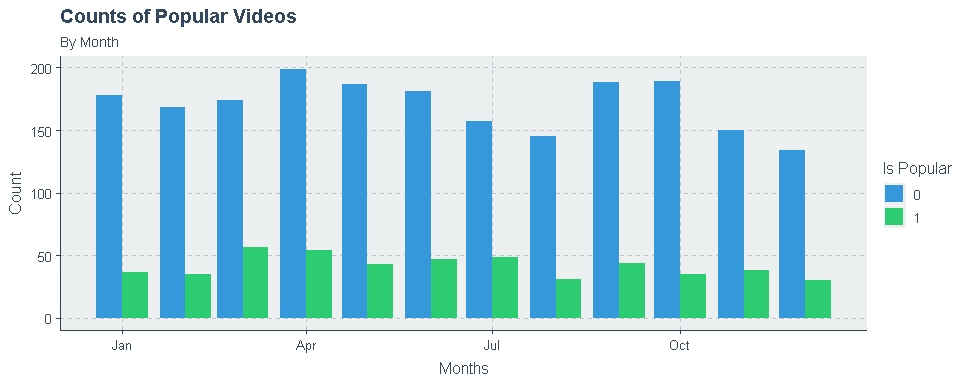
In the graph showing numbers of views by month by year, there is an upward trend of views from 2006 to 2013 and then a downward trend afterwards. In general, it seems that the month with the greatest number of views is May. This may be attributed to university students ending their winter semester in April and finding further educational value from these talks. However, as previously stated, the audience varies so greatly that it is difficult to make any specific inference about these trends without further information. Finally, the downward trend post 2013 may show that TED has declined in popularity and people are less likely to watch their videos.

The count of popular videos by month and by year shows an interesting trend. Post 2013, the number of “not popular” videos has greatly declined, to the point where they seem virtually nonexistent in 2016 and 2017.









We now turn our attention to assessing characteristics of topic types for different videos. Using the data term matrix that was previously created, we evaluated the performance of grouping videos into different k number of topics. To briefly summarize the LDA; it is a generative probabilistic model in which objects are broken into smaller pieces that comprise them. A singular piece for a specific object may be identical or unique amongst other objects in the model. The model’s probabilistic setup is to group the pieces amongst all objects into k number of topic groups. With this, the probability of sampling a specific piece from a topic comprises one of the distributions. Additionally, based on the occurrences of pieces in an object, the probability of sampling a topic type from an object can be derived.

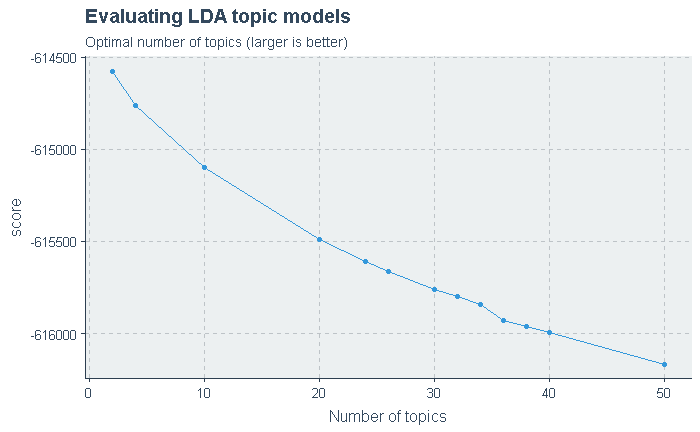
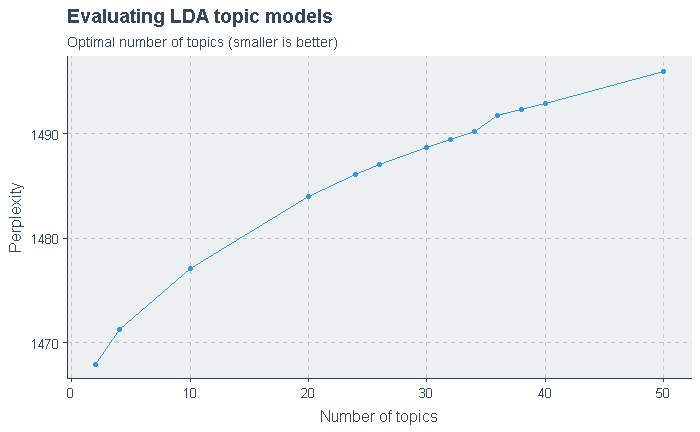
In our situation, the videos represent the objects and the words contained in their title, description, and tags are the pieces they embody. The model aims to estimates two variables: the probability of a word belonging to a topic denoted beta and the probability of a topic belonging to a video denoted gamma. Provided a training set of data, it learns these parameters by using the method of Gibbs sampling. The algorithm goes through each video’s text data and randomly assigns each word within the text data to one of the k topic groups. This forms the initial probability distributions for words in topics and topics in videos. It then goes through each word in each document and evaluates two conditional probabilities. The proportion of the words in the document have been assigned to an individual topic (P(t|v)) and the proportion of that topic to overall videos that come from that specific word (P(w|t). It then reassigns that word to one of the topic groups with probability P(t|v) x P(w|t) (the probability that topic to generate that word). After repeating this process, many times, it eventually reaches a steady state for both probability distributions, after which the process ends, and the resulting estimates are used.

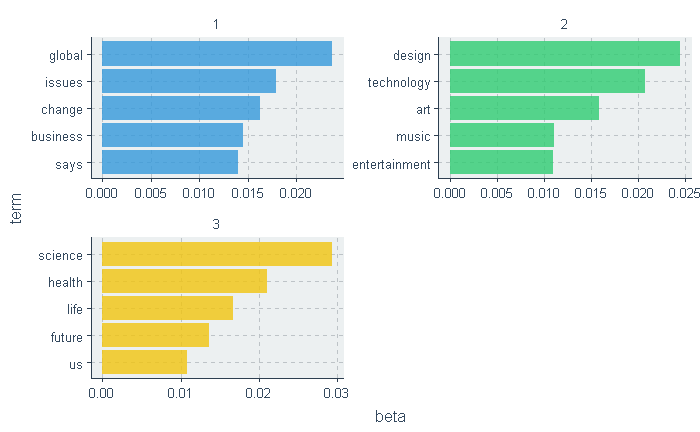
We can then classify each video by comparing the probability of the topic given the video for each topic (gamma) and assigning the topic to the highest probability.

Before the model can be learned, the hyperparameter of k number of groups must be decided. In making this decision evaluated below model results for different levels. Models were compared based on their individual loglikelihoods given the training data and the perplexity of each model. Perplexity, like likelihood, is a measure of how well the model predicts the data. It is, however, a measure of uncertainty in a model. A random variable with perplexity k has the same uncertainty as a k-sided dice.

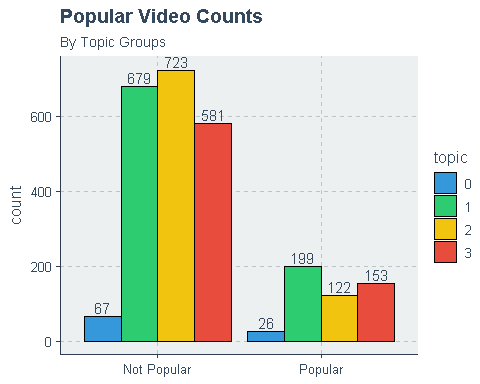
It is important to highlight that, with this type of unsupervised learning, it recommends not solely basing the number of topics off perplexity and log-likelihood but also to asses where the topic types have an intuitive sense to them. For example, whether words are being segmented into topics with an interpretable meaning. With this in mind, we decided on 3 groups. Seen below are the words with the highest probability of belonging to the respective topic group. The model classifies the video into one of these topic groups based on which topic corresponds to the highest probability associated with its respective words and their frequencies. In certain cases, videos have equal probabilities amongst each topic group. In this scenario, we decided to classify it as an arbitrary general topic group which does not have a focused theme.

Here, we see the following topics and their themes: Topic 1 has a theme of culture and art, Topic 2 has a theme of public and international affairs, Topic 3 has a theme of modernism and technology an observation that scored equal probabilities for each topic were assigned a topic value of 0, representing the general topic group.





We then evaluated how popularity differs amongst the respective groups. We noticed that, among the popular videos, Topic 2 (public and international affairs) has the most view count. This makes sense since art/culture and modernism/technology are focused more on groups of people who either work in these sectors or have a personal preference towards them, while public and global issues are shared by a majority of people no matter their personal preferences or employment.



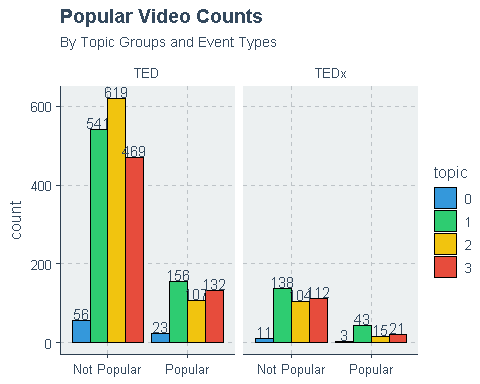
In attempting MANOVA, we see that an assumption of equal covariance was rejected by Wilks’ lambda. This may be because the nature of data is not normal.

## Df Wilks approx F num Df den Df Pr(>F)   
## topic 3 0.96676 7.2111 12 6728.4 0.0000000000002864 \*\*\*  
## Residuals 2546   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

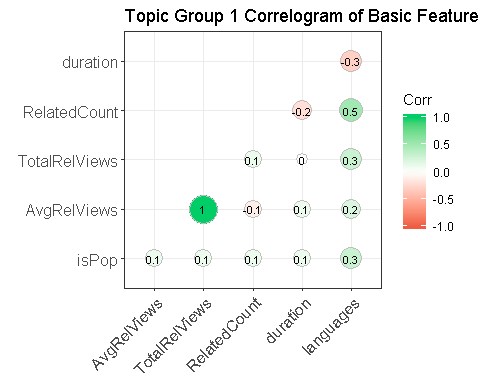
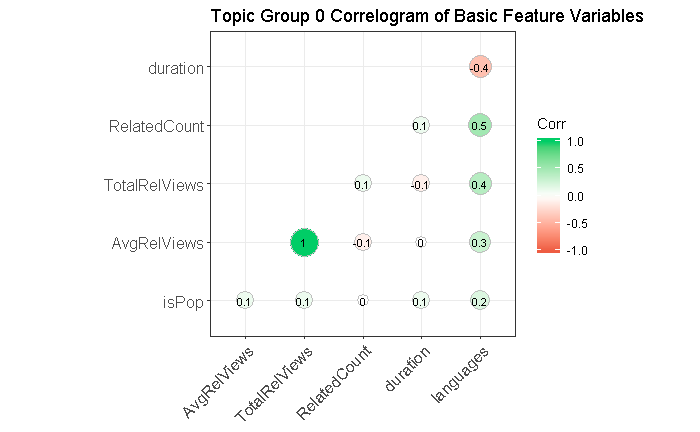
Furthermore, when comparing just the topic groups 1 to 3, the test for equal covariance was still rejected by Wilks’ Lambda. Again, since the nature of data is not normal, it makes sense that the test by Wilks’ Lambda fails.

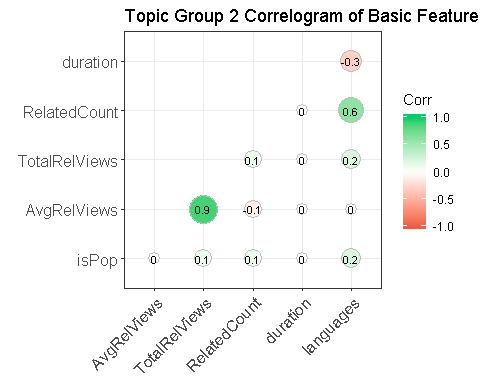
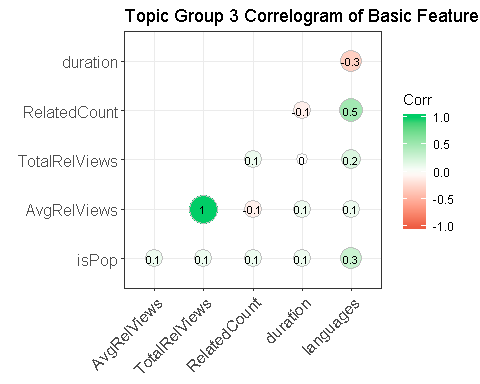
## Df Wilks approx F num Df den Df Pr(>F)   
## topic 2 0.96972 9.4945 8 4902 0.0000000000004064 \*\*\*  
## Residuals 2454   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

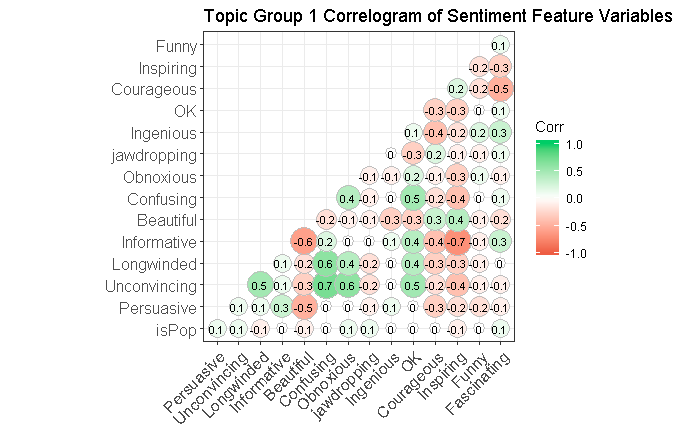
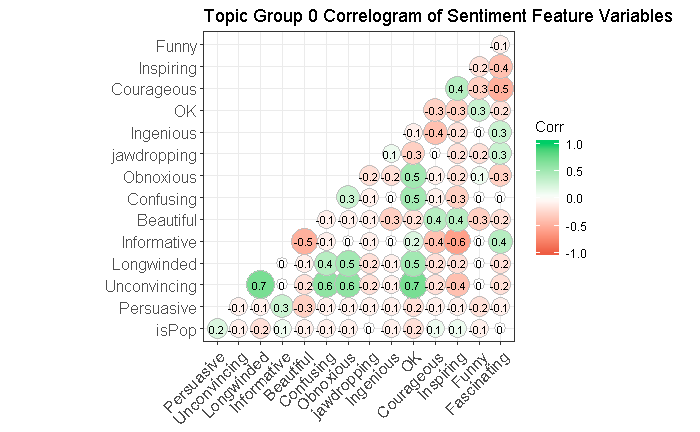
In assessing the popularity of event types amongst the topic groups, we see that Topic 2 has both the most popular TED and TEDx videos. We thought, since the TED and TEDx are intended for different audiences the distribution of popular videos among these two groups might differ as well. However, there does not appear to be any trend reversals when segmenting by event type. Again, it is possible that this is due to the fact that global/public issues are intended to be of interest to a broader audience.

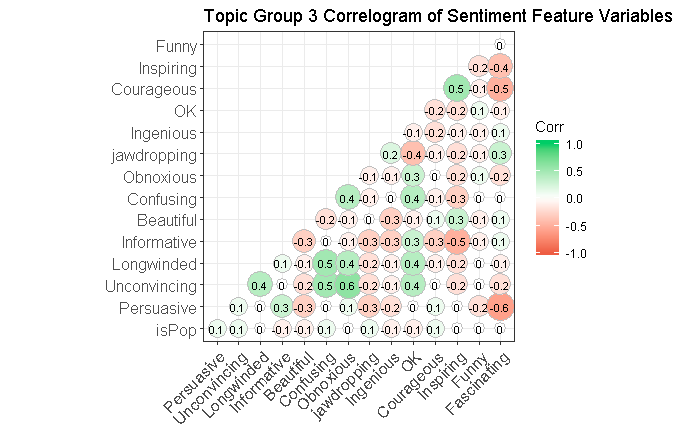
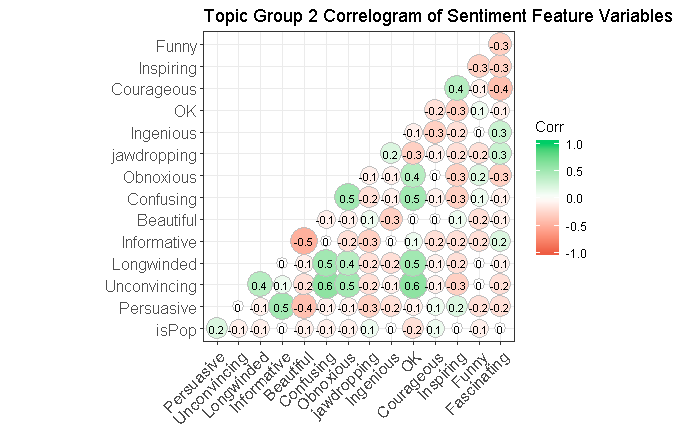


We further evaluated the relationship amongst popularity with feature variables by segments of topic groups. There are not many clear major changes in correlations by topic groups. The number of languages remains the strongest correlation across topics and there are minor differences in correlations amongst the rating ratios. This makes sense since if the video is accessible for more global populations, it will have a higher view count. We observed that, while analyzing the sentiment feature variables, the correlation between comments that express the same degree of evaluation (i.e. “unconvincing” and “confusing” or “inspiring” and “courageous”) is also quite high. Again, we can theorize that the crowd effect is prevalent, since people are more likely to have the same opinion as the majority of the public.





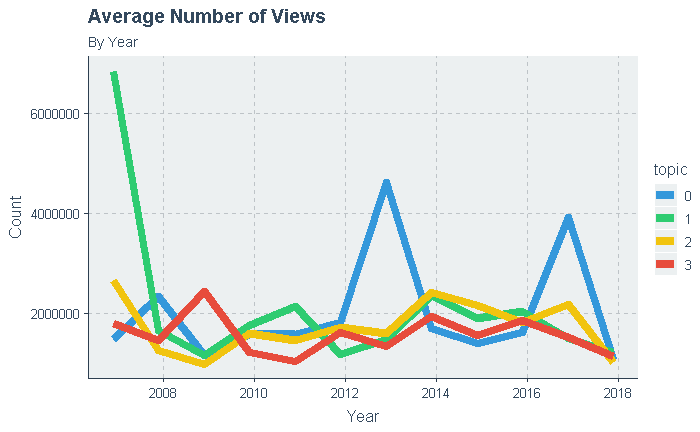
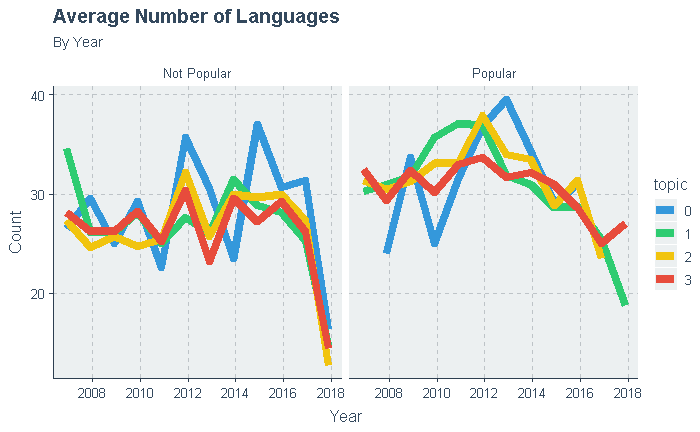


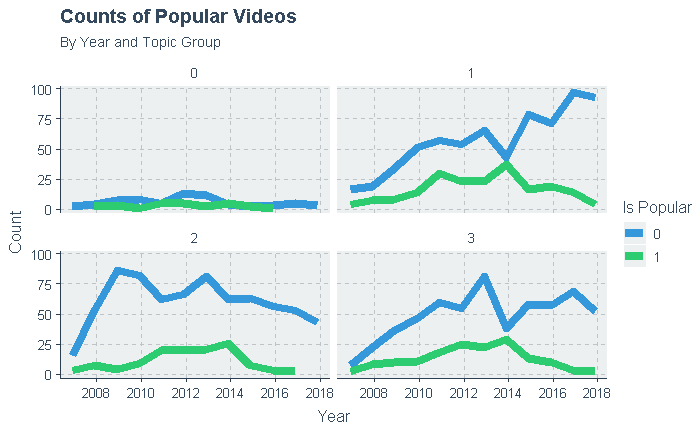
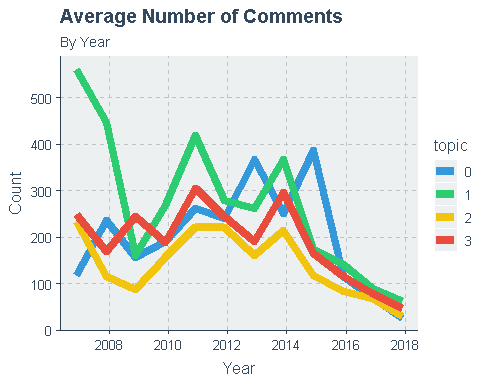


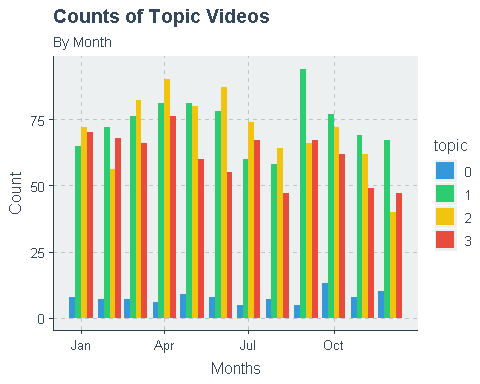
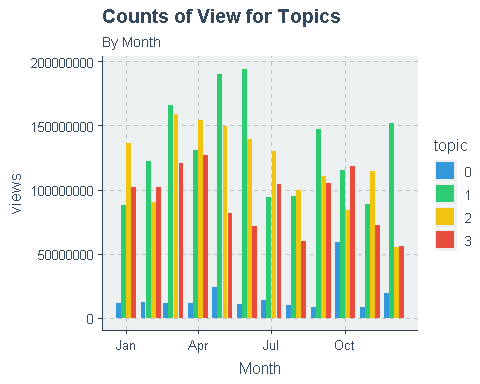
We decided to analyze how the number of views of different topic groups changes over time. As we can see, views for all topic groups are following roughly the same pattern as the popular videos in general. However, it is important to note, that for the general topic group 0 and topic group 3 (art/culture) there were two significant spikes in 2013 and 2014, respectively. Again, we are uncertain of the nature of such behavior, but we can theorize that during these years there were events that spiked peoples’ interests in these topics or there were certain videos in these categories that attracted a lot of attention. The latter theory could be backed up by the graph counts of popular videos by year and topic group, since there is a spike in popular videos in 2014 for group 3.

If we look at the graph that shows us the average number of comments per group topic we can see, that group 2 is generally the most discussed and has spikes in 2008, 2011 and 2014. Group 0, 1, and 3 roughly follow the same pattern, but it is important to note that the largest amount of comments belongs to group 0 in 2014.

We observed that the graphs that portray the counts of videos by topic by month show that group 2 tends to publish more videos from April to June, while group 1 has a spike in videos in September.







To help structure our consideration of predictors of popularity we implemented a forward elimination process with logistic regression. We began by splitting our data in 80% training and 20% testing using stratified random sampling. Our strata sizes were selected proportionally based on the number of popular and unpopular talks available amongst the entire set. We sampled 400 Popular classified videos and 1640 Unpopular classified videos to be used in the training set. The remainder would then be used as a test set to evaluate the trained models.

# Logistic Regression Analysis

We began by training a binomial logistic regression on each individual quantitative and categorical variable. For each model, we checked whether the p-values for all coefficients were less than 0.1. Then the AIK score was reported along with its training and test accuracy. After the first round of variable selection was completed the variable which minimized the AIK was selected to be used in the two variable sequences. We continued this process iteratively. At our 3rd variable selection, p-value significance was then lowered to .075. We continued iteratively selecting the variable that met this criterion and minimized AIK up until a nine-variable model. This is when no additional variable selections had p-values that were significant.

In comparing the top models containing 1 to 8 variables, we noticed that AIK continually decreases, training accuracy continually increases but at a diminishing rate, and training accuracy peaks at the 6 variable model but then begins to decrease with further additional variables.

Since the training accuracy peaks at the model with the 6 variables, in order to avoid overfitting and to make the model simpler, we decided to choose the model with 6 variables.

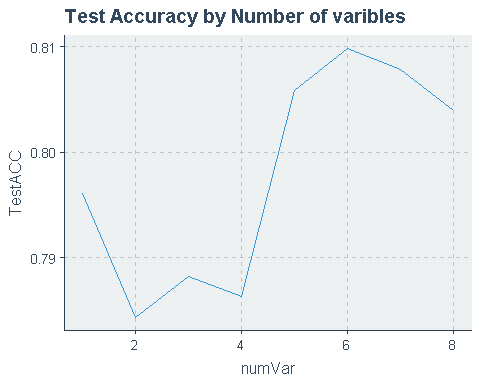
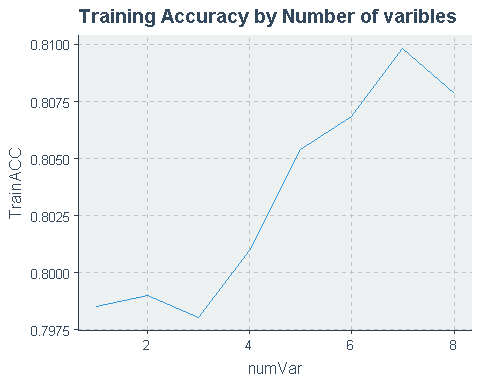
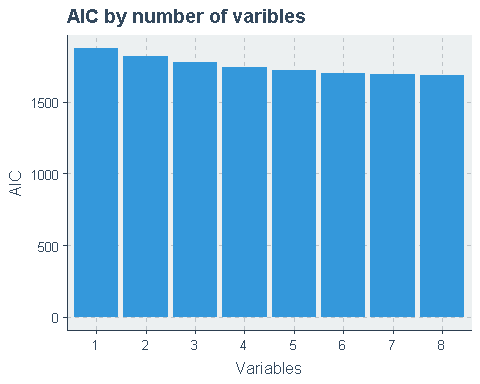
## AIC TrainACC TestACC  
## 1 1869.907 0.7985294 0.7960784  
## 2 1818.953 0.7990196 0.7843137  
## 3 1775.102 0.7980392 0.7882353  
## 4 1739.716 0.8009804 0.7862745  
## 5 1717.235 0.8053922 0.8058824  
## 6 1696.807 0.8068627 0.8098039  
## 7 1688.972 0.8098039 0.8078431  
## 8 1686.238 0.8078431 0.8039216

summary(model)

##   
## Call:  
## glm(formula = fmla, family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2078 -0.6521 -0.4512 -0.1914 3.1743   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.129354 0.418700 -14.639 < 0.0000000000000002 \*\*\*  
## languages 0.100790 0.008168 12.340 < 0.0000000000000002 \*\*\*  
## Persuasive 7.906043 0.956818 8.263 < 0.0000000000000002 \*\*\*  
## OK -16.483585 2.338172 -7.050 0.00000000000179 \*\*\*  
## Unconvincing 11.960625 1.670991 7.158 0.00000000000082 \*\*\*  
## Fascinating 7.301884 1.089275 6.703 0.00000000002036 \*\*\*  
## Courageous 4.814845 1.002973 4.801 0.00000158213508 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2019.3 on 2039 degrees of freedom  
## Residual deviance: 1682.8 on 2033 degrees of freedom  
## AIC: 1696.8  
##

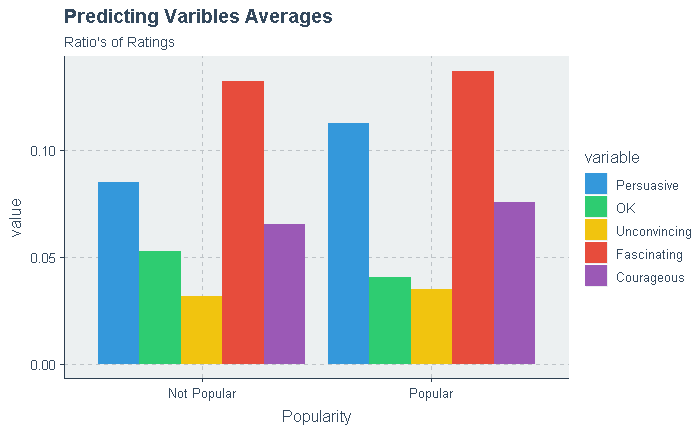
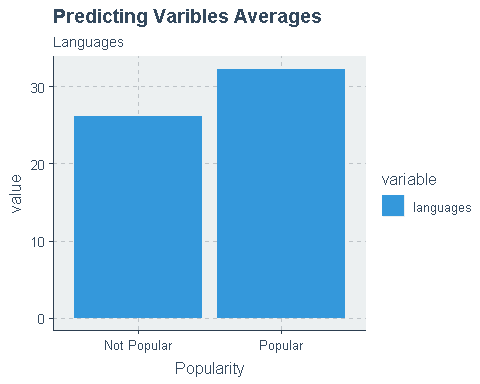
(accuracy(model,train,TED.df))

## [1] 0.8068627 0.8074510

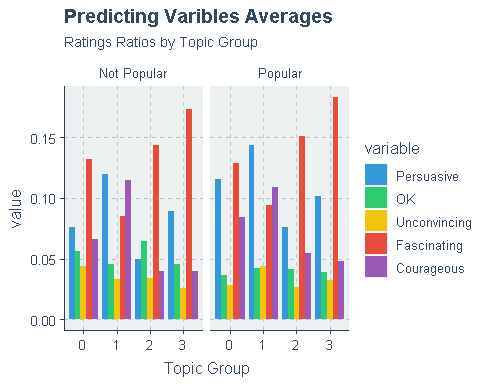
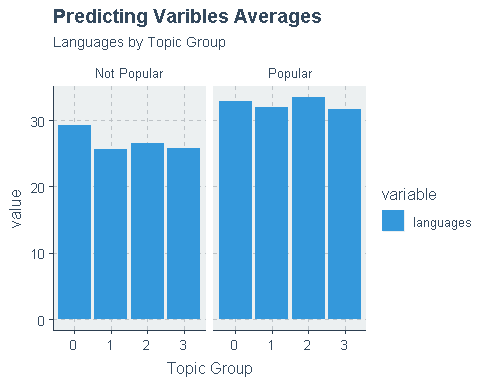


We now wanted to assess more accurately how these six variables characteristics change with respect to popularity, topic groups, and changes in time. The average count of the number of languages is slightly higher amongst popular videos. This agrees with our model’s predictors as it assigns positive but small coefficient of 0.096298 to rating count of number languages. This also supports the notion that expanding the accessibility of potential viewers could increase the popularity of a TED Talk.

In the averages of rating ratios, we see that “Persuasive”, “Courageous”, and “Fascinating” stand out in being larger amongst popular videos. This is in line with our model’s predictors as it assigns positive coefficients to each of these variables. This gives support to the idea that new creators of content will have a greater possibility of having a popular TED Talk if their video can deliver a message containing these themes.



We see that the average ratios for the rating “Fascinating” are highest in topic 3, associated with science and technology, for both popularity categories. This offers some support that viewers of both popular and not popular videos will be more fascinated by this subject matter. Additionally, topic 1, associated with culture and art, has the largest average ratio for “Persuasive”. Overall the averages of ratings do not change drastically between popularity groups when segmented by topic types. This was in line with our predictive model as it did not find topic type have stronger predictive power.



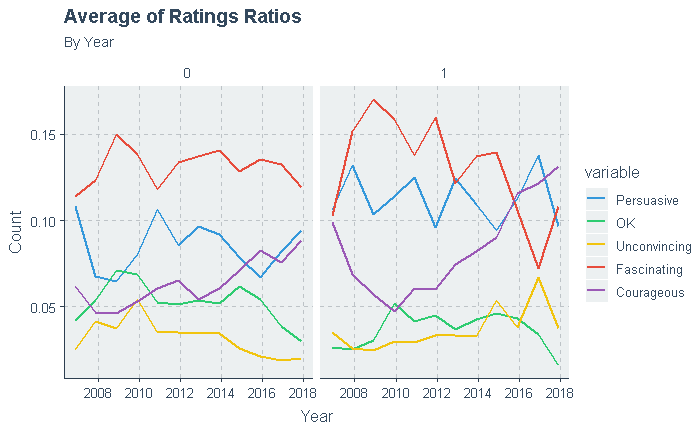
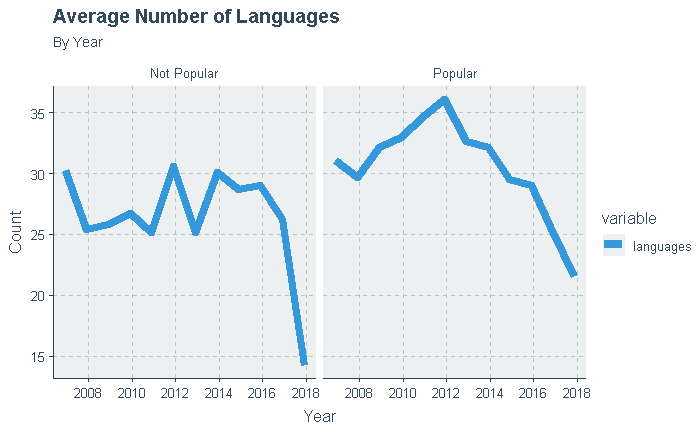
We noticed that the number of languages offered over time for both popular and not popular groups decreases as we move closer to 2017. This makes sense as more recent videos will have had less time to been translated into many languages. However, it is notable that the popular groups have a greater number of languages available for each time period. This supports our model’s predictions as it assigns a positive coefficient to the number of available languages.

In analyzing the ratings ratios over time, we see that the average ratio of “Persuasive” ratings remains consistently higher than that of non-Popular over all time periods. This provides support that, for a TED Talk to be popular, it needs to persuade the viewer of their message they are trying to deliver.

Additionally, the average ratio of “Fascinating” was in the top two average ratios proportions until 2015, where after it fell significantly. Our interpretation of this was that, with more and more quality content being created by different TED Talks over the years, the expectations of viewers have risen over time and viewers are not as easily persuaded as before.

In contrast to this, the average ratio of the rating “Unconvincing” has been trending upwards in recent years. Our model assigns this ratio a positive coefficient. Our intuition behind this is if there is a divide in users’ opinions on whether a TED Talk is persuasive or unconvincing it will instigate debate amongst the comments and increase the user activity in the process.

We see that the average ratio of “Courageous” rating has continuously increased from 2010 onward which supports the idea that a shift in theme preference towards speakers whose topic takes courage to address. We see that the “Okay” rating is consistently lower in the popular category overall years. This is in line with our prediction model which assigns a negative coefficient to ratios of okay. The intuition behind this is that videos that are not able to generate a more extreme opinion from viewers will fail to attract larger viewer activity.



Our findings suggest that future TED Talks will have a greater probability of generating higher views and comments by offering the video in a greater number of languages. Furthermore, our findings suggest that the probability of popularity increases if the speaker is able to either fully persuade their audience or spark a debate amongst viewers who agree and disagree with the message. This will help generate a greater number of views and comments. In recent years, the “Courageous” rating ratio has been associated with popular videos. However, it is recommended that these trends are continued to be watched over the coming year; as with any preference they can change over time.