TED Talk Popularity Analysis

2025-04-03

Authors:

LANGXIN LI (1008831374)
Derrick Li (1009049959)
Qi Sun (1009013050)
Yangxinyue Wang (1008834629)

1. Introduction

1.1 Background & Study Aim

TED Talks is a global platform for communicating innovative ideas.Previous studies have explored elements such as presentation style, duration, thematic content in relation to views and audience engagement (Agarwal, 2021). However, specific factors that contribute to the success of a TED talk remain less clear (Fischer, 2024). The aim of this study is to identify and quantify the attributes that make a TED Talk more popular. This will be done so by modeling popularity as a function using predictors like talk duration, language availability, speaker count, video age, and thematic tags. In addition, we also use nonlinear relationships, polynomial transformations, and interaction terms such as between talk duration and language availability

1.2 Research Question

What factors predict the popularity of TED Talks, and how do they contribute to the popularity score?

- **Hypothesis 1**: TED Talks with longer durations and greater video age are associated with higher popularity scores.
- **Hypothesis 2**: Specific attributes, such as the number of speakers, tag count, and tag contents such as science and global issues, influence the popularity score.

1.3 Dataset Cleaning

1.3.1 Original Dataset Overview

The initial TED Talks dataset includes variables such as metadata (e.g., comments, descriptions, event details, publication dates), content attributes (e.g., duration, language availability, speaker information, ratings, tags), and engagement metrics (e.g., view counts). We only focused on the direct key variables that relate to viewer engagement so:

- **Comments**: The number of first-level comments made on the talk.
- **Duration**: The duration of the talk in seconds.
- **Languages**: The number of languages in which the talk is available.
- **Num_speaker**: The number of speakers featured in the talk.
- **Published_date**: The Unix timestamp for the publication of the talk on TED.com.
- **Ratings**: A string dictionary detailing the ratings given to the talk along with their frequencies.
- **Tags**: The themes or topics associated with the talk.
- **Views**: The total number of views on the talk.

Rationale for Excluding Certain Variables: Variables like **description, event, film_date, related_talks, speaker_occupation, title**, and **url** were excluded because they are either meta-information or require specialized processing such as the natural language processing (NLP) for unstructured text that is outside the scope of the course.

1.4 Cleaning and Reformatting Steps (See Appendix for detail R code)

- ➢ **Initial cleaning:** Removing records with missing values in the key columns, ensuring data integrity for the following analyses.

- ➢ **Date & Time Conversion:** Converting Unix timestamps in *published_date* to human-readable formats, standardizing "*published_date", "video_age"* (current date minus published date).
- ➢ **Tags Processing**: The *tag_count* column and the categorical variables ("*tag_is_technology"* , "*tag_is_science"*, and "*tag_is_global_issues"*) were chosen because they were the most frequent tag among all TED Talks.
- ➢ **Ratings Processing**: The ratings field was reformatted and normalized by converting each video's raw rating counts into proportions of its total ratings. "**Inspiring**" was the most common rating. Indicator variables were created to tell whether each video includes "**Inspiring**" among its top five. This approach controls for variations in the total number of ratings per video.
- ➢ **Target Variables: "***popularity_score"* is defined as "*log(views_per_day) + log(comments)"*, where "*views_per_day"* is the ratio of views to "*video_age"*. This transformation stabilizes variance, reduces skewness, and produces a target variable that is closer to a normal distribution and entirely positive (completed later in Box-Cox transformations).

1.5 Analysis Coverage

1.5.1 Exploratory Data Analysis (EDA):

Use of descriptive statistics and visualizations (e.g. pairs plots, box plots) to examine distributions, assess skewness, and explore relationships among predictors and the target variable.

1.5.2 Model Selection and Validation:

A stepwise AIC procedure was applied to refine the initial full model. The dataset was then split into training (80%) and test (20%) sets for validation, and its predictive performance was evaluated on the test set using RMSE and R-squared.

1.5.3 Post Diagnostic Analysis:

- ➢ **LINE Assumption Checks:**
  The Residuals vs Fitted plot, Q-Q plot, Scale-Location plot, and Residuals vs Leverage plot are used to assess linearity, normality, equal variance, and influence, respectively.

- ➢ **Outlier and Influence Checks:**
  Computing studentized residuals with a Bonferroni-adjusted threshold for outlier detection, calculating hat values (using a 2p/n threshold) for identifying high-leverage observations, and assessing individual influence via Cook's Distance.

2. Exploratory Data Analysis

2.1 Summary of Variables

- **popularity_score:** A composite metric computed as log(views_per_day) + log(comments), capturing video popularity.

- **duration:** The length of the TED Talk in seconds.
- **languages:** The number of languages in which the talk is available.
- **num_speaker:** The number of speakers featured in the talk.
- **video_age:** The age of the video in days, calculated as the difference between the current date and the published date.
- **tag_count:** The number of tags associated with the talk.
- **Is_rating_t5_Inspiring:** A binary indicator (0/1) denoting whether “Inspiring” is among the top 5 ratings for the video.
- **tag_is_technology:** A binary indicator (0/1) showing whether the talk is associated with the “technology” tag.
- **tag_is_science:** A binary indicator (0/1) indicating whether the talk is associated with the “science” tag.
- **tag_is_global_issues:** A binary indicator (0/1) indicating whether the talk is associated with the “global issues” tag.

```
##  popularity_score    duration        languages      num_speaker
##  Min.   : 5.354   Min.   : 135.0   Min.   : 0.00   Min.   :1.000
##  1st Qu.: 9.423   1st Qu.: 597.8   1st Qu.:23.00   1st Qu.:1.000
##  Median :10.271   Median : 862.0   Median :28.00   Median :1.000
##  Mean   :10.381   Mean   : 839.8   Mean   :27.29   Mean   :1.025
##  3rd Qu.:11.252   3rd Qu.:1054.0   3rd Qu.:32.00   3rd Qu.:1.000
##  Max.   :17.262   Max.   :5256.0   Max.   :72.00   Max.   :5.000
##    video_age      tag_count      Is_rating_t5_Inspiring tag_is_technology
##  Min.   :2751   Min.   : 1.000   Min.   :0.0000         Min.   :0.0000
##  1st Qu.:3675   1st Qu.: 5.000   1st Qu.:0.0000         1st Qu.:0.0000
##  Median :4621   Median : 6.000   Median :0.0000         Median :0.0000
##  Mean   :4602   Mean   : 7.674   Mean   :0.3593         Mean   :0.3039
##  3rd Qu.:5475   3rd Qu.: 9.000   3rd Qu.:1.0000         3rd Qu.:1.0000
##  Max.   :6856   Max.   :32.000   Max.   :1.0000         Max.   :1.0000
##  tag_is_science   tag_is_global_issues
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.2597   Mean   :0.1952
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

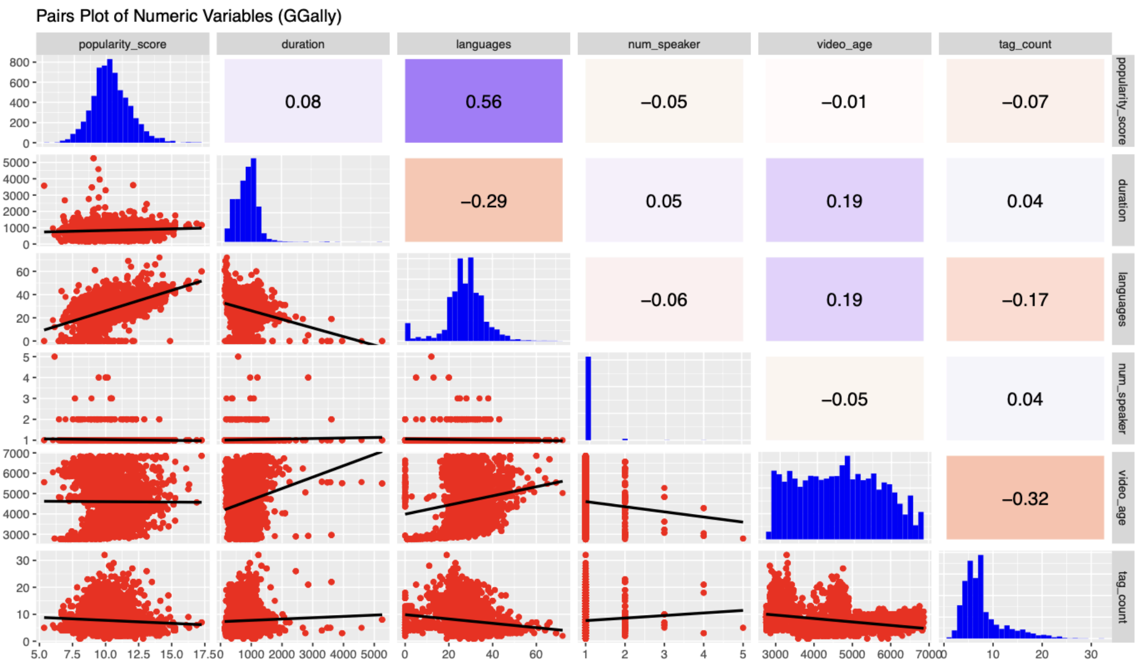
2.2 Quantitative Variables:

Integrated Interpretation of the Pairs Plot:

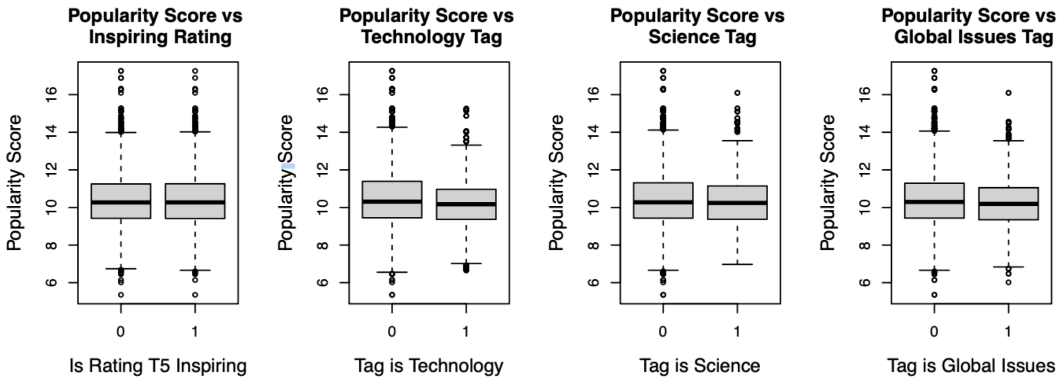
The pairs plot provides a comprehensive view of both the individual distributions and the relationships among variables. Overall, the **popularity score** is fairly symmetric, indicating that most videos cluster around a moderate level of popularity. In contrast, variables such as **duration, video age**, and **tag count** show right-skewed distributions, with most values on the lower end but a few high outliers.

Among the relationships, the **number of languages** stands out as the strongest predictor of popularity, with a moderate positive correlation (≈ **0.56**). This suggests that videos available in more languages tend to be more popular.

Looking at the explanatory variables themselves, there is a moderate negative correlation (≈ **-0.32**) between **video age and tag count**, indicating that older videos tend to have fewer tags. Additionally, a slight negative association (≈ **-0.29**) is observed between **languages and duration**, implying that videos offered in more languages may be slightly shorter.



2.3 Qualitative Variables

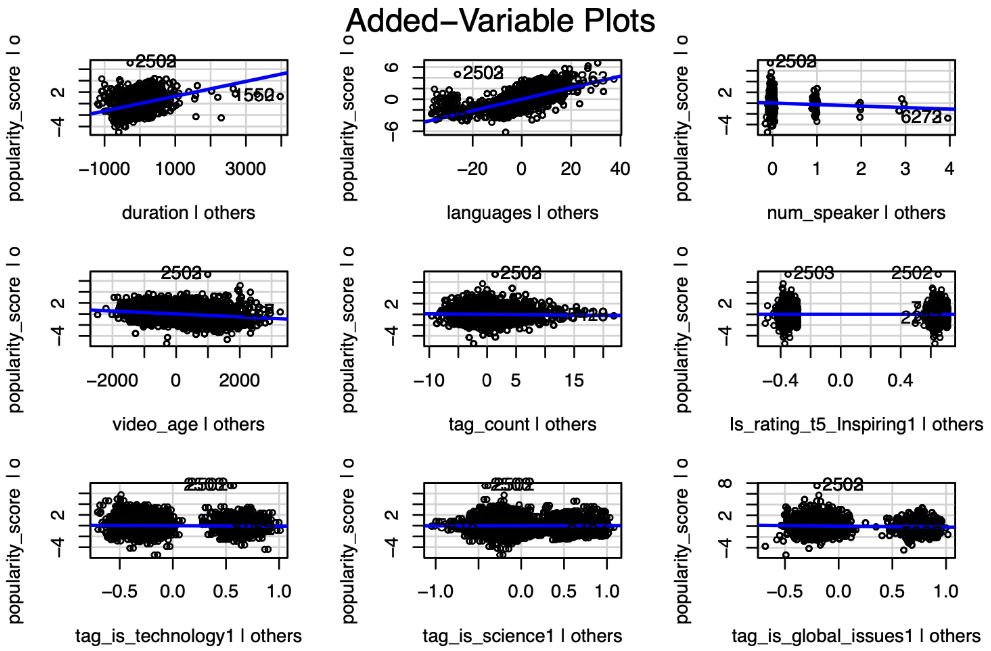


- **Inspiring Rating**: The “Yes” group shows a slightly higher median popularity score than the “No” group, though both distributions overlap considerably. This suggests that having a Is_Inspiring_rating may be linked to somewhat higher popularity, but the effect does not appear very large.
- **Technology Tag**: Videos tagged with “Technology” have a median popularity score similar to those without the tag. There is a modest spread in both groups, and a few outliers extend well above the median, indicating that while some “Technology” videos are highly popular, many remain at moderate popularity levels.
- **Science Tag**: The difference in medians between “Yes” and “No” for the Science tag is relatively small. As with the Technology tag, both groups display wide variability, with overlapping interquartile ranges and some high outliers.
- **Global Issues Tag**: The two categories (“Yes” vs. “No”) show overlapping distributions and similar medians. Although there are a few outliers in the “Yes” group, the data suggests that including a Global Issues tag alone does not strongly differentiate a talk’s popularity.

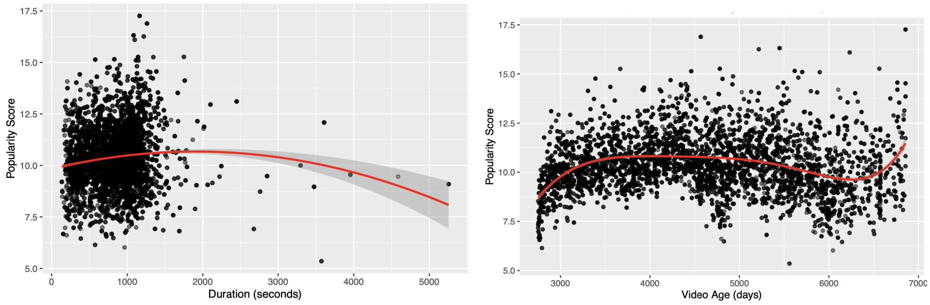
3: Fit Initial Full Model (Main Effects Only)

```
full_model <- lm( popularity_score ~ duration + languages + num_speaker + video_age + tag_count +
tag_is_technology_ +  tag_is_science + tag_is_global_issues,data = ted_data)
```

4. Check for Linearity, Curvature, and Interaction Effects



Curved Relationship between Duration and Popularity (Left)
Curved Relationship between Video Age and Popularity (Right)



From the plot above, we notice:

- An overall curved trend, specifically for duration|Others, which have many outliers on its right. This implies a curved or non-monotonic relationship, so **higher degree polynomial terms** are needed.
- num_speaker|Others: we notice the plot is flat and sparse, with most points lying on 0 and 1. This implies it is a **weak relationship** and **predicator**.
- language|Others: we notice strong positive linear trends, which implies that as the number of languages increases, popularity scores tend to increase as well.
- video_age|Others: we notice a decreasing U trend so **higher degree polynomial** terms are needed.

- tag_count|Others: we notice a very flat slope, suggesting that this has little to no contribution as a predictor.
- tag_is_technology|Others, tag_is_science|Others, tag_is_global_issues|Others: we notice that the slopes are all flat as well. This suggests that they might have **little to none individual contribution as a predictor**, but when used together, it could be meaningful.

  Hence, we retain them all for the next step.

```
##  RESET test
##
## data:  full_model
## RESET = 420.75, df1 = 2, df2 = 6372, p-value < 2.2e-16
```

The reset value is 420 which is a bit too large and p-value is relatively small, which means we would reject the null hypothesis. This means this current model is lacking important **interactions and non-linear relationships and polynomial terms**.

4.1 Findings:

AV plots and residual vs. fitted plots for key predictors, particularly for duration and video_age, showed clear signs of curvature, a U-shape, suggesting a high-degree polynomial. This is also supported by the RESET test.

Based on domain knowledge and model diagnostics, we suspected that the effects of some predictors might vary depending on the level of language diversity (**languages**). To explore this, we included interaction terms: **duration:languages** and **num_speaker:languages** were added to examine whether the influence of a video's duration and the number of speakers on **popularity_score** changes across different language contexts. Also, **video_age:languages** to test whether the effect of a video's age on popularity depends on how linguistically diverse the content is.

4.2 Choices of Degree:

Duration

The initial guess for duration is of degrees 2. We fit the following model. When increasing the degree of video_age from 2 to 3, both R^2 and $R^2_{Adjusted}$ only increase only by less than 0.01. Therefore we keep it to degree 2.

Video Age

The initial guess for video age is of degrees 2. Notice that by increasing the degree of video_age from 2 to 6, both R^2 and $R^2_{Adjusted}$ improved by about more than 0.03. Therefore we keep it to degree 6. However, if it increases from 6 to 7, both R^2 and $R^2_{Adjusted}$ do not improve much, only by about 0.0003. Thus, we keep it to degree 6.

Therefore, their related interaction terms are updated to **poly(duration, 2):languages** and **poly(video_age, 6):languages**

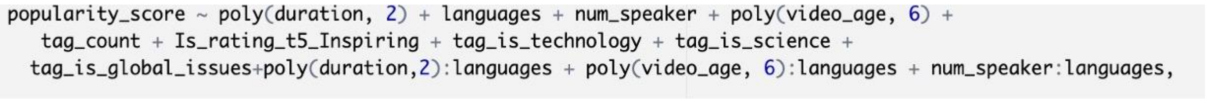
(k, n)	(2,2)	(3,2)
R^2	0.468	0.4707
R^2_{adj}	0.467	0.4692
AIC	19149.95	19149.95
BIC	19149.95	19140.5
PRESS	7378.552	7347.228

(k, n)	(2,2)	(2,3)	…(2,5)	(2,6)	(2, 7)
R^2	0.468	0.480	0.4904	0.497	0.497
R^2_{adj}	0.467	0.479	0.489	0.495	0.495
AIC	19149.95	19140.5	19022.73	18933.34	18867.7
BIC	19149.95	19022.73	18933.34	18867.7	18881
PRESS	7378.552	7218.66	7092.572	7004.898	7006.914

Conclusion:

- Polynomial terms (e.g., **poly(duration, 2)**) were added when the AV plots showed curvature in the relationship between duration and popularity_score.
- Interaction terms (e.g., **poly(duration, 2):languages**, **poly(video_age, 6):languages**, and **num_speaker:languages**) were included because we confirmed through testing that the effects of these predictors depend on the level of language diversity.

Our full model is defined below:



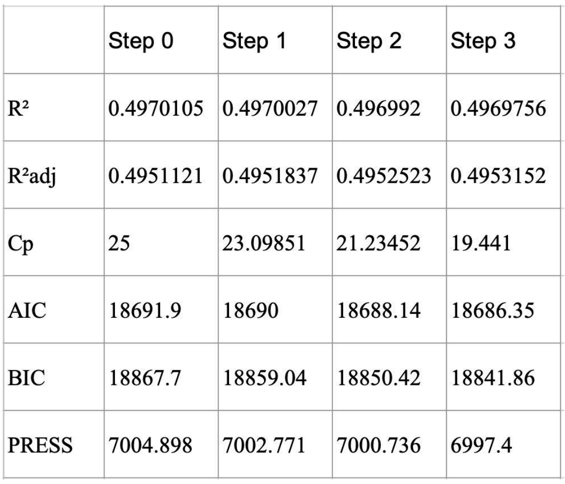
```
popularity_score ~ poly(duration, 2) + languages + num_speaker + poly(video_age, 6) +
   tag_count + Is_rating_t5_Inspiring + tag_is_technology + tag_is_science +
  tag_is_global_issues+poly(duration,2):languages + poly(video_age, 6):languages + num_speaker:languages,
```

5. Model Selection

After getting the full model, we use stepwise AIC regression to find our step AIC model:

```
step_model_aic <- step(final_model, direction = "both")
```

	Step 1	Step 2	Step 3
poly(duration, 2)	-	-	-
languages	-	-	-
num_speaker	-	-	-
poly(video_age, 6)	-	-	-
tag_count	-	-	-
Is_rating_t5_Inspiring	NA	NA	NA
tag_is_technology	-	NA	NA
tag_is_science	-	-	-
tag_is_global_issues	-	-	-
poly(duration, 2):languages	-	-	-
poly(video_age, 6):languages	-	-	-
num_speaker:languages	-	-	NA



	Step 0	Step 1	Step 2	Step 3
R^2	0.4970105	0.4970027	0.496992	0.4969756
R^2adj	0.4951121	0.4951837	0.4952523	0.4953152
Cp	25	23.09851	21.23452	19.441
AIC	18691.9	18690	18688.14	18686.35
BIC	18867.7	18859.04	18850.42	18841.86
PRESS	7004.898	7002.771	7000.736	6997.4

During the process, **Is_rating_t5_Inspiring, tag_is_technology and languages:num_speaker** are removed from the full model.

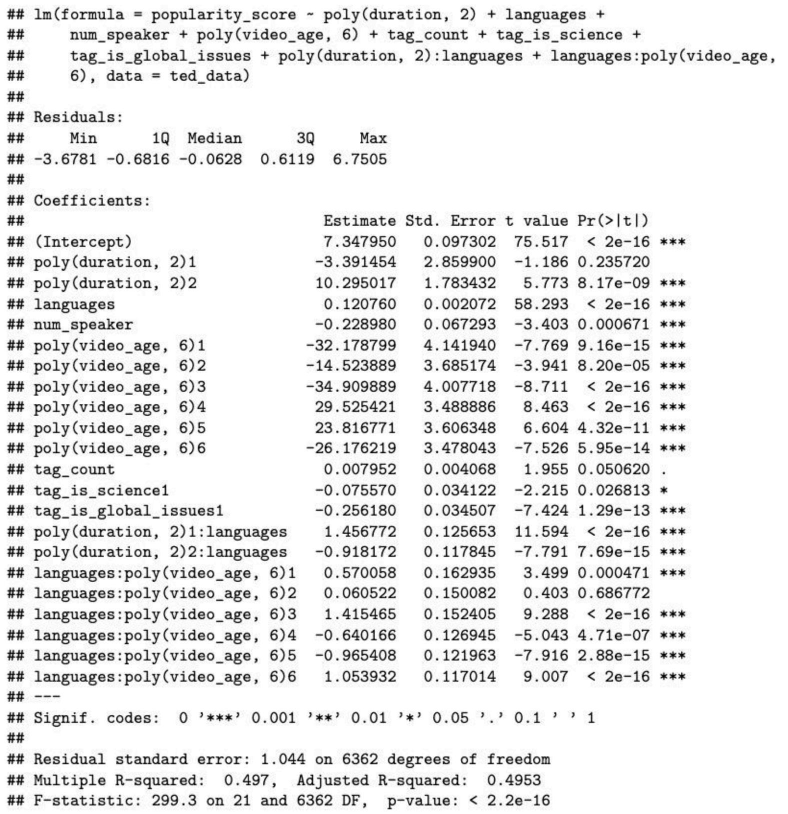
The stepwise regression process effectively simplified the original model by removing predictors and interaction terms that contributed minimally to explaining variations in TED Talk popularity.

The final model (Step 3) retained key variables, including duration, languages, number of speakers, video age, tag count, thematic tags (“science” and “global issues”), and selected interactions.

Model simplification resulted in negligible loss of explanatory power (Adjusted R² decreased marginally from 0.4951 to 0.4953), while significantly improving predictive performance criteria, including reduced values of Mallows’ Cp (from 25 to 19.44), AIC (18691.9 to 18686.35), BIC (18867.7 to 18841.86), and PRESS (7004.9 to 6997.4).

These improvements demonstrate that the refined model is statistically superior, balancing explanatory robustness with enhanced interpretability and predictive accuracy.

```
summary(step_model_aic),
```



```
## lm(formula = popularity_score ~ poly(duration, 2) + languages +
##     num_speaker + poly(video_age, 6) + tag_count + tag_is_science +
##     tag_is_global_issues + poly(duration, 2):languages + languages:poly(video_age,
##     6), data = ted_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6781 -0.6816 -0.0628  0.6119  6.7505
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.347950   0.097302  75.517  < 2e-16 ***
## poly(duration, 2)1            -3.391454   2.859900  -1.186 0.235720
## poly(duration, 2)2            10.295017   1.783432   5.773 8.17e-09 ***
## languages                      0.120760   0.002072  58.293  < 2e-16 ***
## num_speaker                   -0.228980   0.067293  -3.403 0.000671 ***
## poly(video_age, 6)1          -32.178799   4.141940  -7.769 9.16e-15 ***
## poly(video_age, 6)2          -14.523889   3.685174  -3.941 8.20e-05 ***
## poly(video_age, 6)3          -34.909889   4.007718  -8.711  < 2e-16 ***
## poly(video_age, 6)4           29.525421   3.488886   8.463  < 2e-16 ***
## poly(video_age, 6)5           23.816771   3.606348   6.604 4.32e-11 ***
## poly(video_age, 6)6          -26.176219   3.478043  -7.526 5.95e-14 ***
## tag_count                      0.007952   0.004068   1.955 0.050620 .
## tag_is_science1               -0.075570   0.034122  -2.215 0.026813 *
## tag_is_global_issues1         -0.256180   0.034507  -7.424 1.29e-13 ***
## poly(duration, 2)1:languages   1.456772   0.125653  11.594  < 2e-16 ***
## poly(duration, 2)2:languages  -0.918172   0.117845  -7.791 7.69e-15 ***
## languages:poly(video_age, 6)1  0.570058   0.162935   3.499 0.000471 ***
## languages:poly(video_age, 6)2  0.060522   0.150082   0.403 0.686772
## languages:poly(video_age, 6)3  1.415465   0.152405   9.288  < 2e-16 ***
## languages:poly(video_age, 6)4 -0.640166   0.126945  -5.043 4.71e-07 ***
## languages:poly(video_age, 6)5 -0.965408   0.121963  -7.916 2.88e-15 ***
## languages:poly(video_age, 6)6  1.053932   0.117014   9.007  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 6362 degrees of freedom
## Multiple R-squared:  0.497,  Adjusted R-squared:  0.4953
## F-statistic: 299.3 on 21 and 6362 DF,  p-value: < 2.2e-16
```

5.1 Key Predictors:

- β_0 (Intercept = 7.34795): Estimated intercept of the regression line. It represents expected popularity_score when all predictors (all polynomial components and interaction terms) are at their baseline or reference values (ie: the reference language and for average values of duration and video_age as represented in the transformed space).

poly(duration, 2):

- β_1 (poly(duration,2)1 = –3.39145): Its p-value (0.236) indicates that this particular component is not statistically significant but interpretable in terms of raw duration. But it contributes, along with poly(duration,2) to modeling the nonlinear effect of duration popularity_score.
- β_2 (poly(duration,2)2 = 10.29502): Its p-value is significant (p < 2e–16) and it captures curvature in the effect of duration on popularity_score. So it indicates that the relationship between duration and popularity_score is nonlinear.

languages, β_3 : For one-unit change in the languages variable, the popularity_score is estimated to increase by 0.12076, holding all other factors constant. This effect is highly significant (p <2e–16).
num_speaker, β_4 : Holding all other variables constant, each additional unit in num_speaker is associated with a decrease in popularity_score by about 0.229. And it is statistically significant (p = 0.000671).

tag_is_science:

- β_{12}: Holding other factors constant, talks tagged as science have a popularity score that is 0.07557 lower than those not tagged as science. This effect is significant(p = 0.02681).
- β_{13}: Holding other factors constant, talks tagged as global issues have a popularity score 0.25618 points lower than those without. This effect is statistically significant (p = 1.29e–13).

tag_count, β_{11}: For each additional tag, the popularity_score increases by 0.008, holding all other predictors constant. Its p-value (0.05062) is marginally significant.

poly(video_age, 6)1-6:

- β_5 to β_{10} poly(video_age,6) components): These six coefficients (–32.17880, –14.52389,–34.90989, 29.52542, 23.81677, –26.17622) together model the nonlinear relationship between video_age and popularity_score. all has p value < 1e–4 indicates that video_age has a strong effect on popularity_score when modeled via a 6th-order polynomial.

languages:poly(duration, 2):

➢ β_{14}: First component of duration polynomial changes by 1.45677 in language category when moving from one language group to another. It is highly significant (p < 2e–16).

➢ β_{15}: The effect of the second component of the duration polynomial is 0.91817 lower for the specified language group compared to the reference. Significance is (p = 7.69e–15) so it is a differential nonlinear effect of duration by language.

languages:poly(video_age, 6)1-6:

➢ β_{16}: A one-unit increase is associated with an additional 0.57006 increase in popularity_score for the given language group (p = 0.000471).

➢ β_{17}: A one-unit increase is associated with an additional 0.060522 increase in popularity_score for the given language group. Not statistically significant (p =0.68677).

➢ β_{18}: A significant positive interaction (p < 2e–16), so the effect of video_age on popularity_score is 1.41547 greater in specified language groups.

➢ β_{19}: Negative coefficient (p = 4.71e–07) implies a drop of 0.64017 in that language group.

➢ β_{20}: A significant negative interaction (p = 2.88e– 15) indicates a decrease of 0.96541 in the effect of this component on popularity_score for the language group.

➢ β_{21}: Positive and significant coefficient (p <2e–16) shows the effect of the sixth component is 1.05393 higher for the given language group.

5.3 Model Comparison:

	Main Effect Model	Full Model	Step AIC Model
R_squared	0.4147304	0.4970105	0.4969756
Adjusted_R_squared	0.4139041	0.4951121	0.4953152
Mallows' Cp	1035.218	25	19.441
AIC	18841.86	18691.9	18686.35
BIC	19703.48	18867.7	18841.86
PRESS	8095.527	7004.898	6997.4

Notice that our predictive power increased from 0.41473 to 0.4969756, and the Adjusted R square increases, AIC, BIC, PRESS, Cp decrease. Hence we chose the Step AIC Model shown below:

$$\begin{aligned}\mathrm{lm}(\text{formula} = \text{popularity\_score} \sim\ & \text{poly}(\text{duration},\ 2) + \text{languages} + \text{num\_speaker} + \text{poly}(\text{video\_age},\ 6)\\ & + \text{tag\_count} + \text{tag\_is\_science} + \text{tag\_is\_global\_issues}\\ & + \text{poly}(\text{duration},\ 2) : \text{languages} + \text{languages} : \text{poly}(\text{video\_age},\ 6))\end{aligned}$$

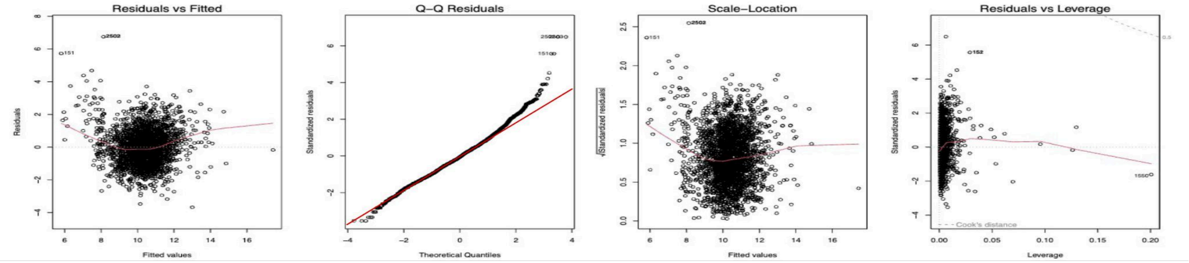
5.4 Data Validation:

RMSE	MAE	R-squared	MSPR	MSE_F
1.038	0.823	0.52	1.077	1.089

The model was trained on 80% of the data validation, and the RMSE is 1.038, and the R-squared is 0.52, which means that the model explains approximately 52% of the variance in the popularity score. For variable exclusion, variables such as Is_rating_t5_Inspiring and languages:num_speaker were excluded during the stepwise process because their p-values exceeded 0.05.

6. Model Diagnostics

6.1 LINE Assumption Checks



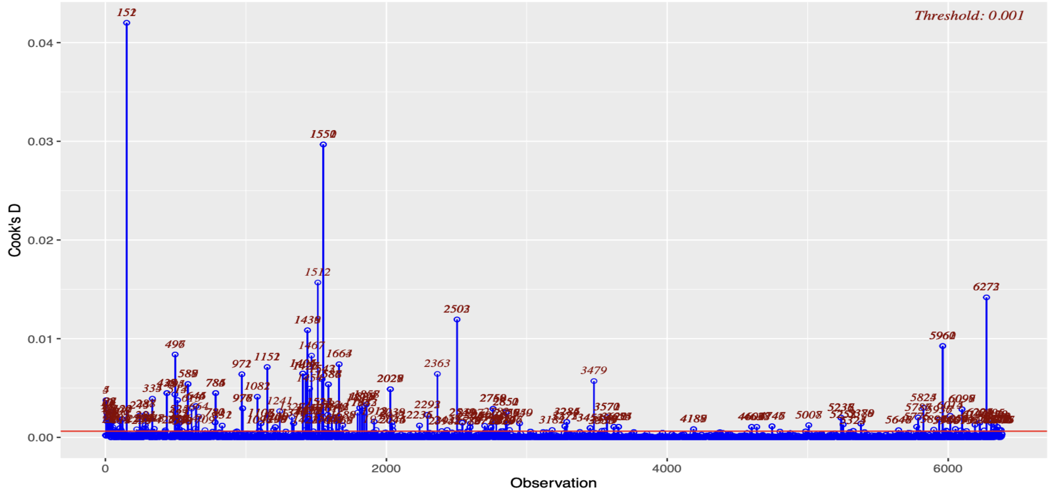
The diagnostic analysis of regression assumptions (LINE) indicates that the fitted regression model satisfies essential criteria, though minor deviations exist. Residual plots show generally random scattering around zero, indicating reasonable linearity, despite minor curvature that suggests slight nonlinearity. The normality of residuals is largely upheld, though moderate deviations at the extreme tails suggest a slightly heavier-tailed distribution. Homoscedasticity (equal variance) appears broadly satisfied, with minimal evidence of increasing variance at higher fitted values.

The Residuals vs Leverage plot indicates a small number of potentially influential observations; however, their limited Cook's distances suggest a minimal impact on overall model performance. Collectively, these diagnostics indicate the model is robust and reliable. There is no severe violations compromising the current model's validity were identified. Consider using Box-Cox transformation and Weighted Least Squares for potential non-constant variance.

6.2 Outlier, Leverage, and Influence Diagnostics

	Outlier Diagnostics	Leverage Diagnostics	Influence Diagnostics	Total num of Observations
Count	5	534	366 (by Cook's)	6384

By computing studentized residuals, the diagnostic analysis identified 5 outliers out of 6384 total observations, indicating only a small fraction of data points differ significantly from model predictions. Using hat values, 534 observations exhibited high leverage, suggesting unusual predictor values or combinations. Influence diagnostics further revealed 366 potentially influential observations, meaning these points could notably affect model results. Given the large dataset size, however, the overall impact of these points remains moderate. The next graph investigates the influential data points further, using Cook's Distance.



Cook’s Distance identified **366 out of 6,273** that exceed the threshold ≈ **0.000637**. This shows the model is table for most data points. Observations points like #1512, #152, #1550, #6273, and were flagged as extreme influential data points. These points are not removed at this stage unless a clear data error is identified.

7. Conclusion

7.1 Key findings:

Our analysis shows the current model only captures about half of what makes a TED talk popular, which means there's plenty of room to improve. For example, when a talk is available in more languages, it tends to attract a wider audience (positive impact of 0.12). On the other hand, talks with more speakers seem to be less popular (negative effect of -0.22). When it comes to talk duration, it turns out that "longer is better" does not always apply, the effect is complex and nonlinear. Similarly, the impact of the video age is also complex, nonlinear, but fluctuates.

7.2 Limitations:

However, the study has several limitations. First, the choice of model and data type restrict the scope of possible findings, and there could be additional variables that are missing that could enhance predictive power. Here, the LINE assumptions are only partially met. Applying Weighted Least Squares improves predictive power to 51.15%. Furthermore, features like tags or metadata that are stored as arrays or dictionaries are not suited for traditional models like linear regression. Additionally, poorly engineered features can fail to capture relevant information and introduce bias, particularly when some categories are rare or misrepresented.

7.3 Future Research:

In the future, we plan to expand the data dimensions, perform sentiment analysis using NLP on speeches, and integrate audience interaction data such as likes and shares. Also, we can apply advanced modeling techniques such as the hierarchical methods. These steps are expected to further reduce the impact of outliers and better capture the interactions between variables and accuracy when it comes to predicting the popularity of TED talks.

References

Agarwal, Vaishali & Tyagi, Vastav & Shivangi, S.. (2021). Investigating the Factors that Contributes most to the Virality of a Social Media Video advertisement. SMS Journal of Entrepreneurship & Innovation. 7. https://doi.org/10.21844/smsjei.v7i01.28728

Fischer, Olivia & Jeitziner, Loris & Wulff, Dirk. (2024). Effect in science communication: A data-driven analysis of TED Talks on YouTube. Humanities and Social Sciences Communications. 11.https://doi.org/10.1057/s41599-023-02247-z

Optional

We used Power BI for demonstration, please see the document on Quercus named “optional_power_bi.pbix”.