TED TALK
POPULARITY ANALYSIS

LANGXIN LI, QI SUN, YANGXINYUE WANG, DERRICK LI

TED

RESEARCH CONTEXT

What factors predict the popularity of TED Talks?
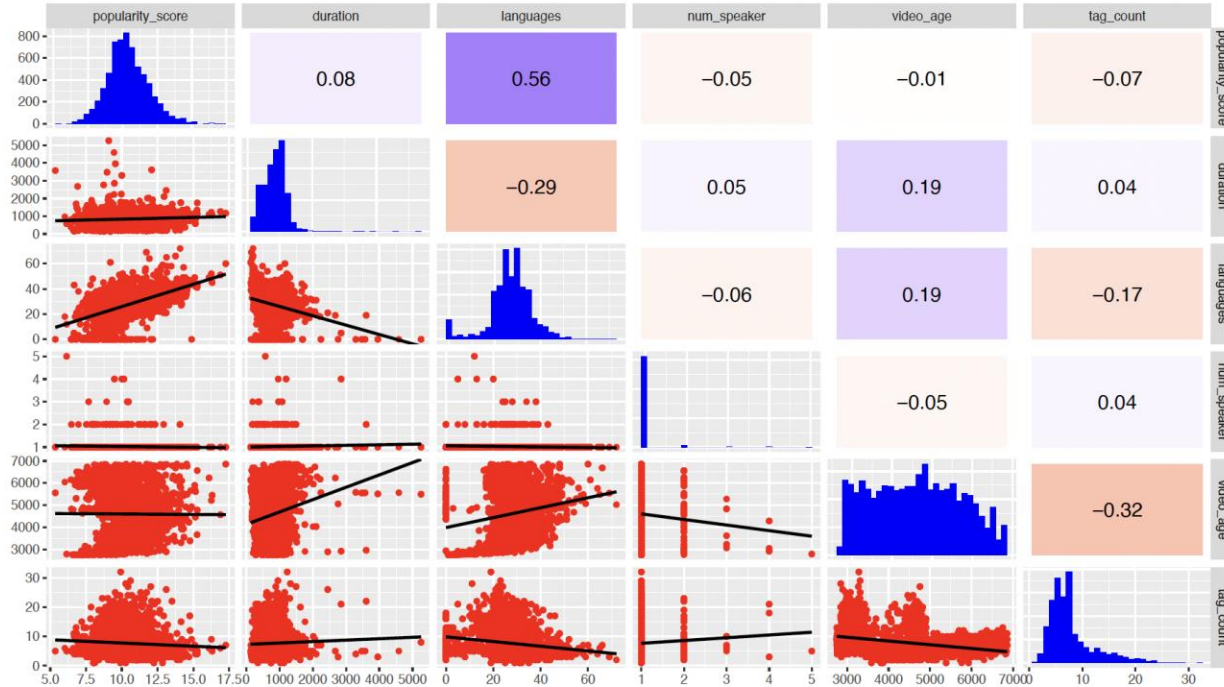
- duration?
- topics?
- number of languages?
- publish date?
- or others ...... ?

TED

SUMMARY OF VARIABLES

| 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	A composite metric computed as log(views_per_day) + log(comments), cap turing video popularity

TED

VARIABLE CORRELATION ANALYSIS

Pairs Plot of Numeric Variables (GGally)

POPULARITY SCORE VS LANGUAGES

Popularity Score vs Languages
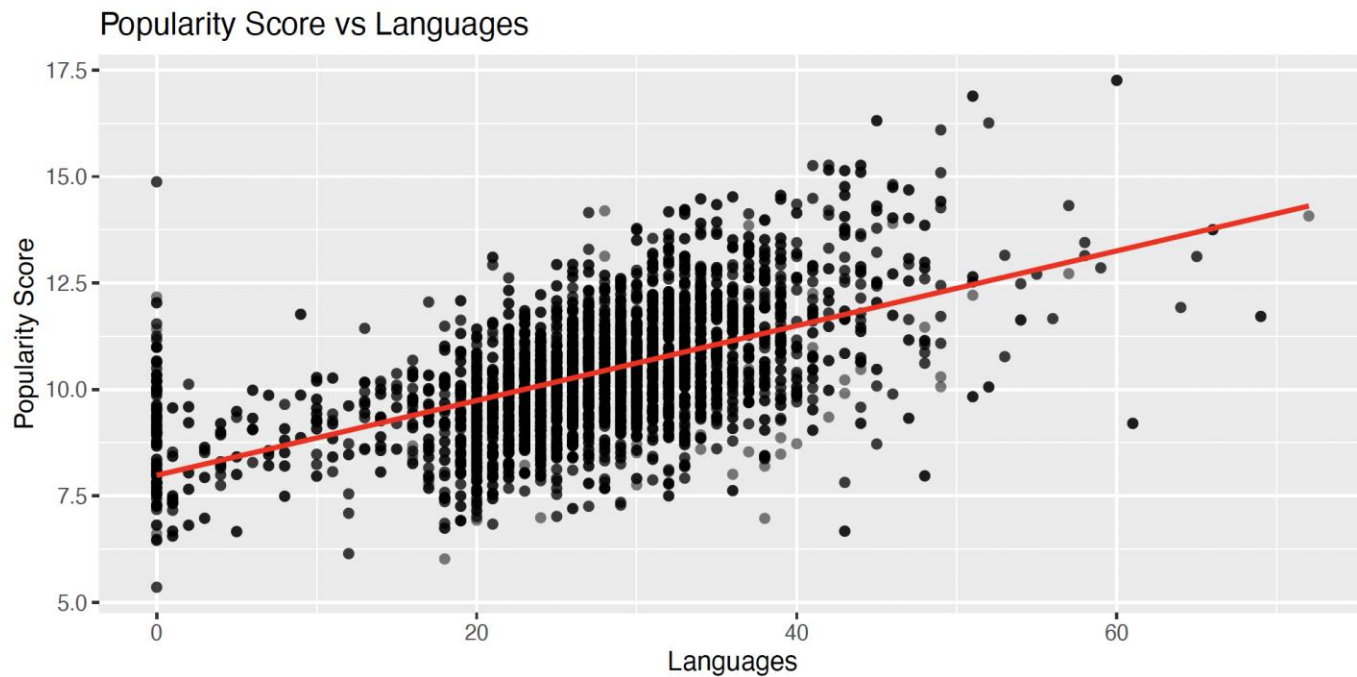
POPULARITY SCORE VS TAGS AND RATINGS
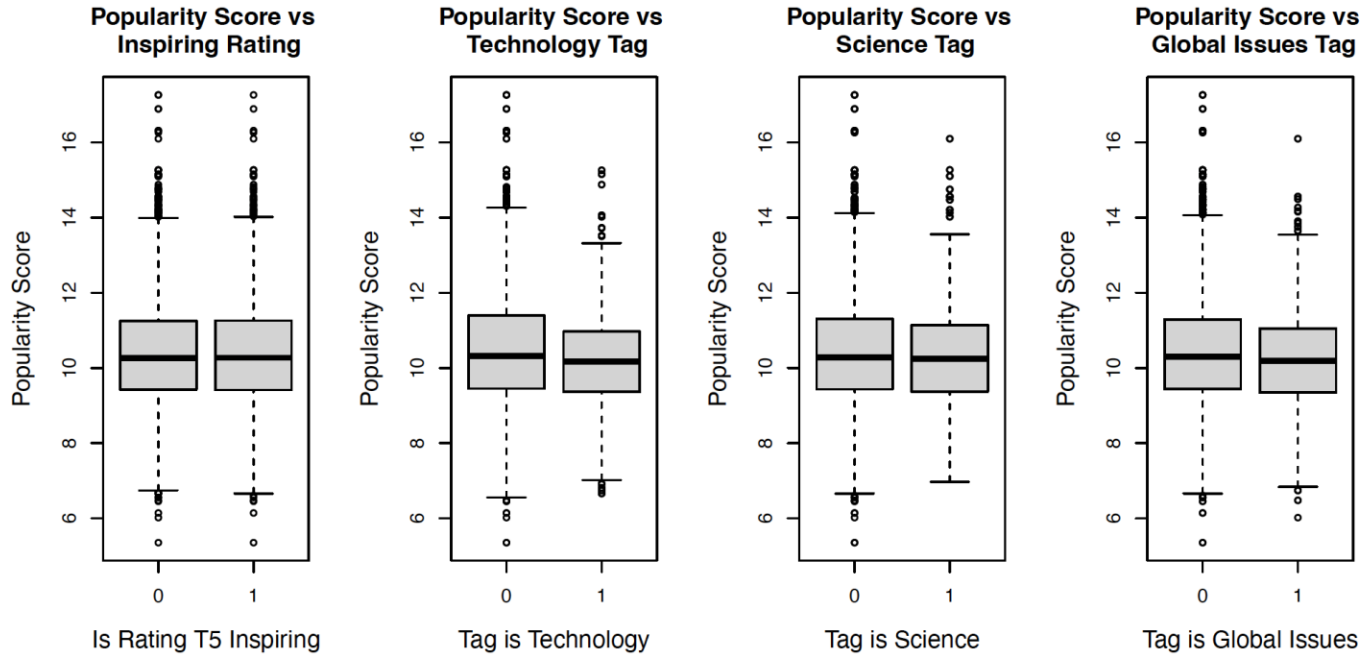
MAIN EFFECT MODEL

```
popularity_score ~ duration + languages + num_speaker + video_age + tag_count + Is_rating_t5_Inspiring +
  tag_is_technology + tag_is_science + tag_is_global_issues
```

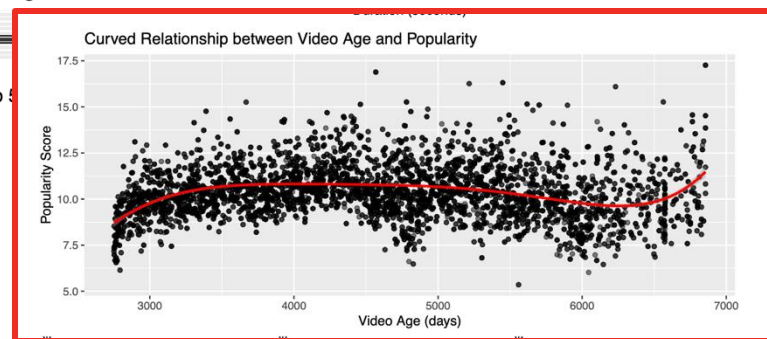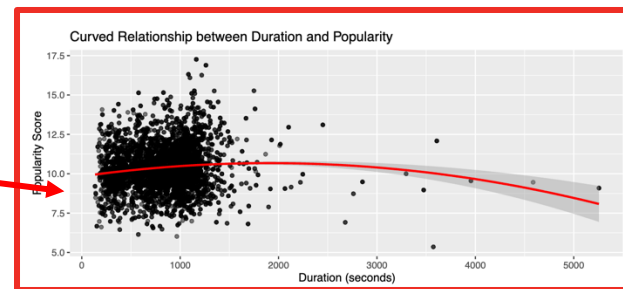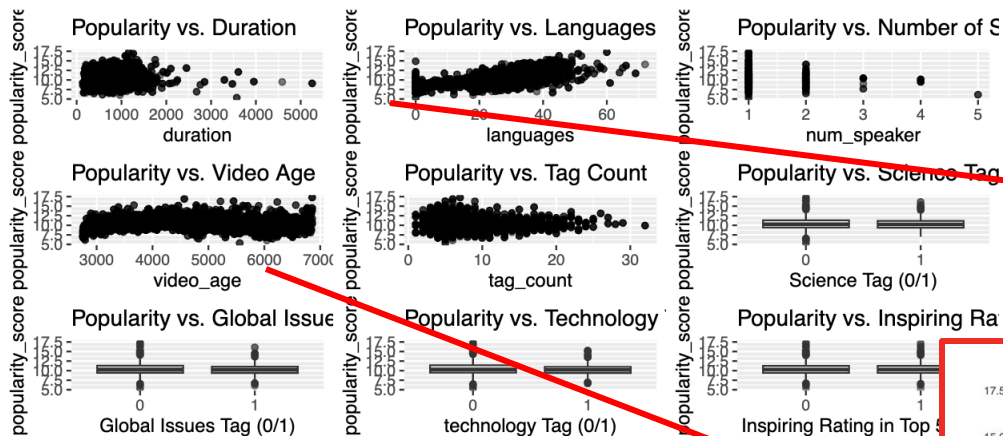R-squared: 0.4147304
Adjusted R-squared: 0.4139041
AIC: 18841.86
BIC: 19703.48
Cp 1035.218
PRESS: 8095.527

CHECK FOR LINEARITY & CURVATURE

```
popularity_score ~ duration + languages + num_speaker + video_age + tag_count + Is_rating_t5_Inspiring +
    tag_is_technology + tag_is_science + tag_is_global_issues
```

INTERACTION TERMS

Based on domain knowledge, add interaction between
- Duration and languages
- Video age and languages
- Number of speaker and languages

Full model:

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

How k and n are defined?

TED

POLYNOMIAL TERMS

Full model:

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

FULL MODEL

Full model:

```
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,
```

R^2	0.497
R^2_{adj}	0.495
Cp	25
AIC	18933.34
BIC	18867.7
PRESS	7004.898

STEPWISE REGRESSION BASED ON AIC:

	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

STEPWISE REGRESSION BASED ON AIC:

	ΔR^2	ΔR^2_{adj}	ΔCp	ΔAIC	ΔBIC	$\Delta PRESS$
Step 0	0.4970105	0.4951121	25	18691.9	18867.7	7004.898
Step 1	0.4970027	0.4951837	23.09851	18690	18859.04	7002.771
Step 2	0.496992	0.4952523	21.23452	18688.14	18850.42	7000.736
Step 3	0.4969756	0.4953152	19.441	18686.35	18841.86	6997.4

Adjusted R Squared Increases

AIC, BIC, PRESS, Cp Decreases

STEP-AIC MODEL

	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

Prediction power increased from 0.4139041 to 0.4953152

AIC FINAL MODEL

```
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)
```

AIC FINAL MODEL

	Estimate	Pr(>|t|)
(Intercept)	7.347950	< 2e-16 ***
poly(duration, 2)1	-3.391454	0.235720
poly(duration, 2)2	10.295017	8.17e-09 ***
languages	0.120760	< 2e-16 ***
num_speaker	-0.228980	0.000671 ***
poly(video_age, 6)1	-32.178799	9.16e-15 ***
poly(video_age, 6)2	-14.523889	8.20e-05 ***
poly(video_age, 6)3	-34.909889	< 2e-16 ***
poly(video_age, 6)4	29.525421	< 2e-16 ***
poly(video_age, 6)5	23.816771	4.32e-11 ***
poly(video_age, 6)6	-26.176219	5.95e-14 ***
tag_count	0.007952	0.050620 .
tag_is_science1	-0.075570	0.026813 *
tag_is_global_issues1	-0.256180	1.29e-13 ***
poly(duration, 2)1:languages	1.456772	< 2e-16 ***
poly(duration, 2)2:languages	-0.918172	7.69e-15 ***
languages:poly(video_age, 6)1	0.570058	0.000471 ***
languages:poly(video_age, 6)2	0.060522	0.686772
languages:poly(video_age, 6)3	1.415465	< 2e-16 ***
languages:poly(video_age, 6)4	-0.640166	4.71e-07 ***
languages:poly(video_age, 6)5	-0.965408	2.88e-15 ***
languages:poly(video_age, 6)6	1.053932	< 2e-16 ***

MODEL DIAGNOSTICS

Cook's D Chart
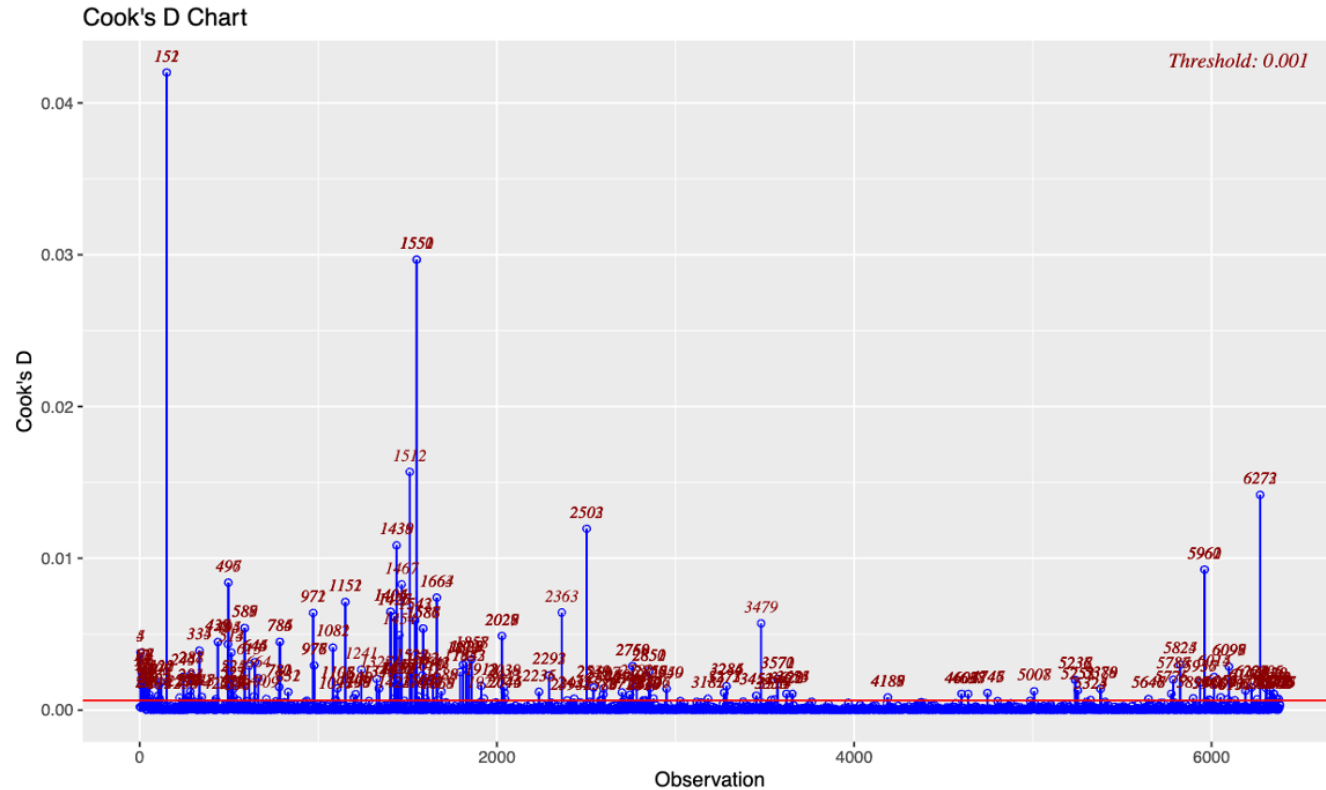
LINE ASSUMPTIONS CHECK
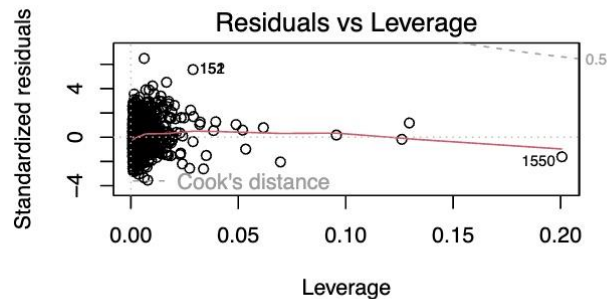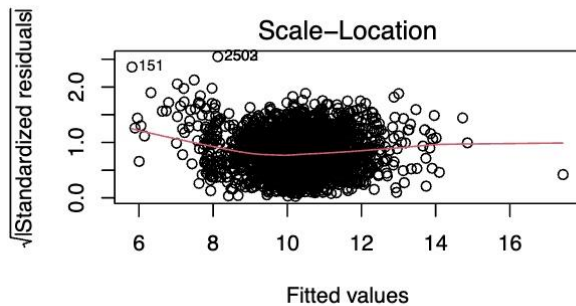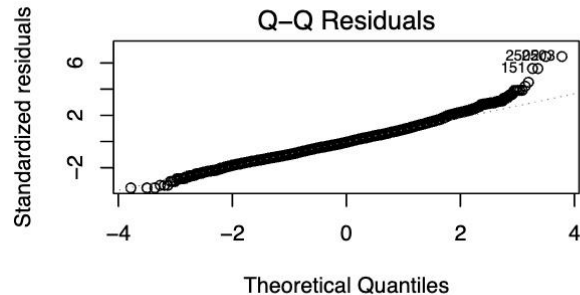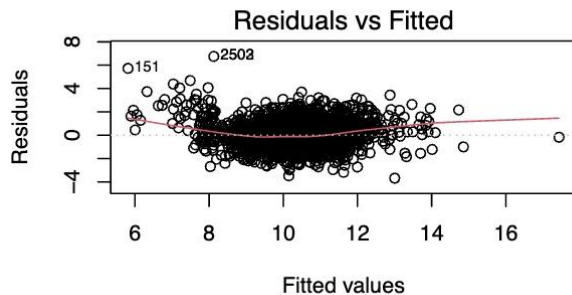
DATA VALIDATION

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

CONCLUSION

IMPACT

- Only 49.7% prediction power

- Languages: 0.12

- Num of speakers: -0.22

- Duration: strong nonlinear effect, rather than "longer is better".

- Video age: complex, nonlinear, but fluctuates

LIMITATIONS

- Model choice vs. data type

- Some potential variables

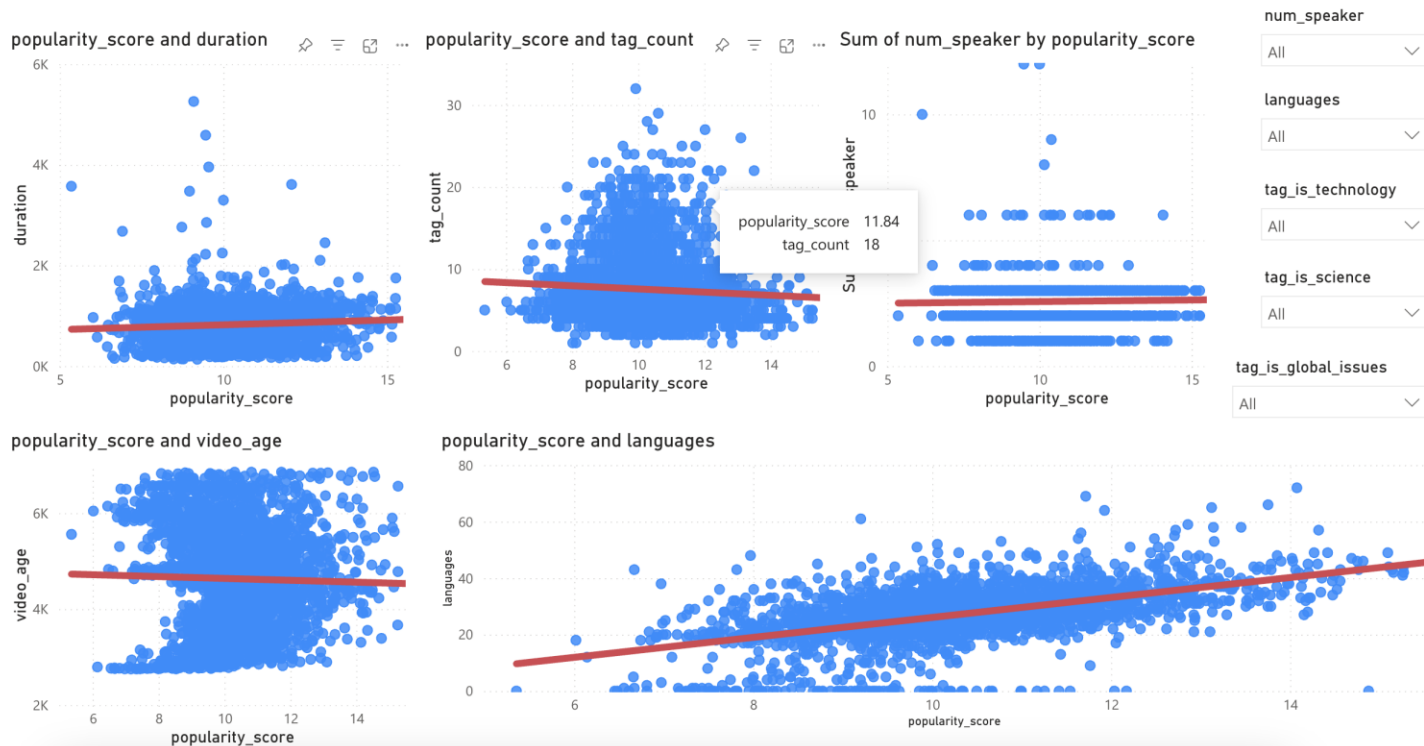- LINE assumptions could be improved by using WLS (51.15% prediction power)

FUTURE

- Expand the data dimensions

- Alternative Popularity Metrics

- Advanced Modeling Techniques

- …

TED

THANK YOU

Linear, linear, linear, linear

Linear, linear, uh, uh-huh, uh-huh

Regression, Regression, Regression, Regression

Regression, Regression, uh, uh-huh, uh-huh

F-test, T-test holds for our slope, yeah

We measure alpha, MLE (uh-huh, uh-huh)

we test we test linear regression

Come tell me what I should reject, oh-oh-oh