Capstone Project

TED TALK VIEWS PREDICTION

By – RAHUL SINGH WALDIA

TED

Well, I’ve analyzed a dataset of 2550 ted talks to get some answers for this question. I explored which of the available variables of a given talk, such as the number of comments, number of languages translated, duration of the talk,

number of tags, or day it was published online– are a strong predictor of its popularity, measured in number of views

After analyzing and regressing views over the other available variables in the dataset, some interesting associations were revealed.

What are features of a popular talk with high number of views?

High number of comments (naturally).

Translations in many languages (also, naturally).

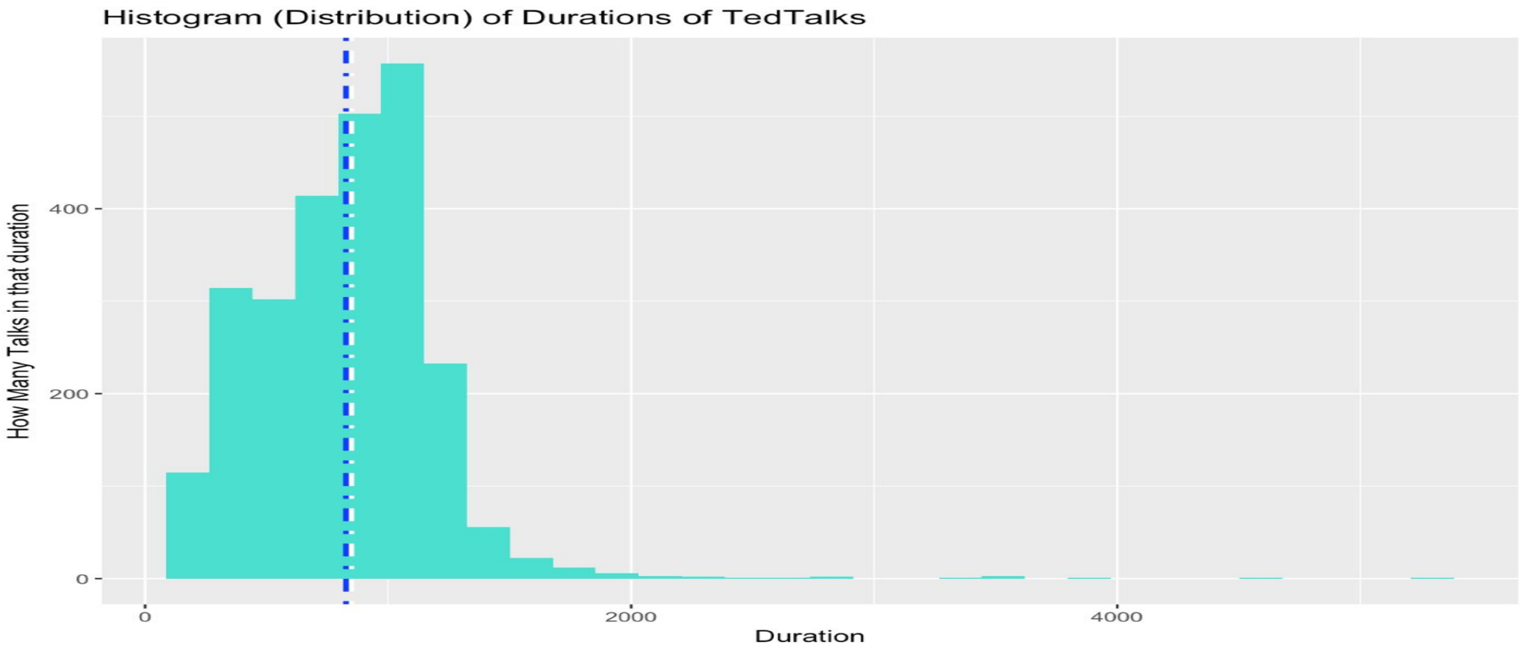
The combination of many comments *and* many translation languages yields a much higher view count than expected by each of them alone.

It *shouldn’t be too short*. Duration didn’t have a big effect at all, but if any, it was positive for longer talks. The most popular talks were between 8–18 minutes.

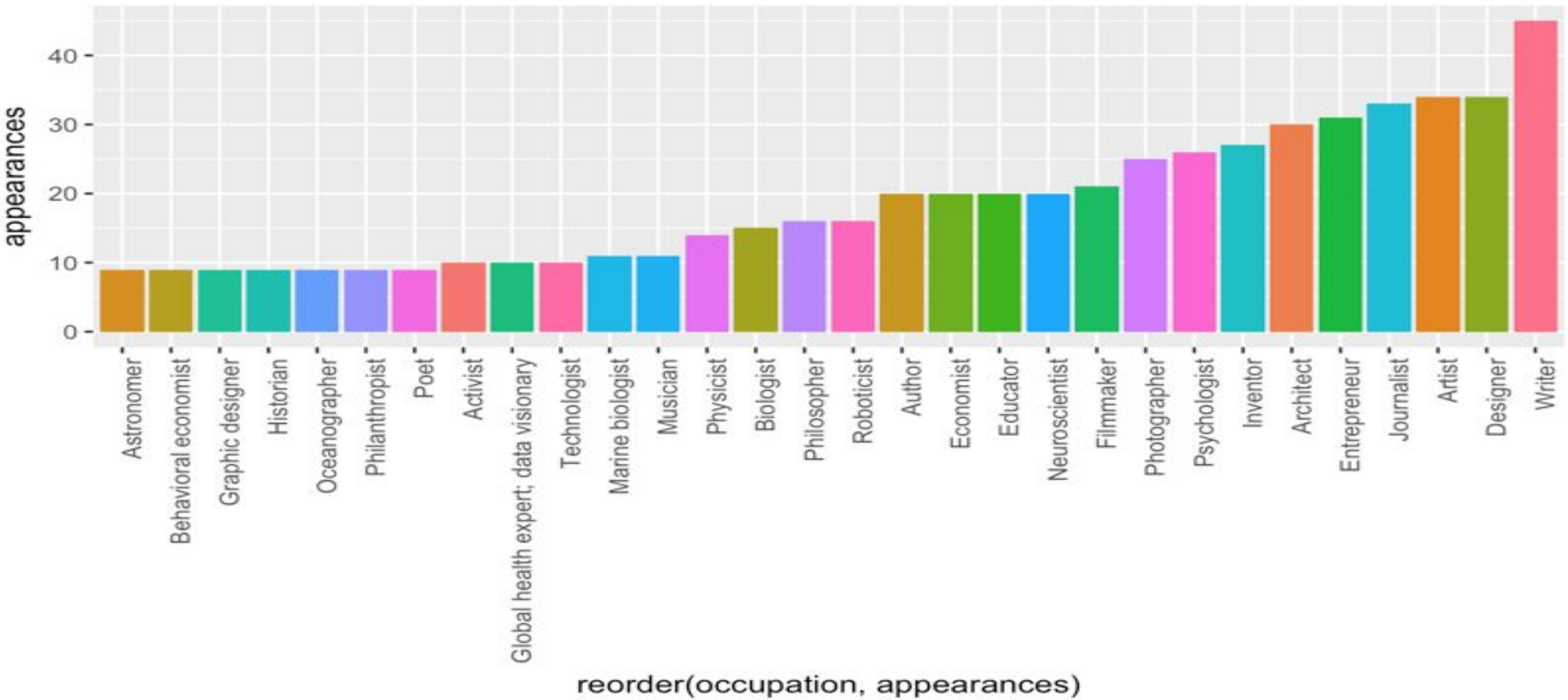
Higher number of tags, ideally between 3–8.

It would be uploaded on a weekday, preferably a Friday!

You may see some funky occupations yielding much higher than average views for their talks, such as: Neuroanatomist, Quiet revolutionary, Lie detector, Model, beatboxer, Vulnerability researcher, or Zen Priest. This isn’t representative, but they did yield the highest number of votes combined (which is an unfair game, but hey).

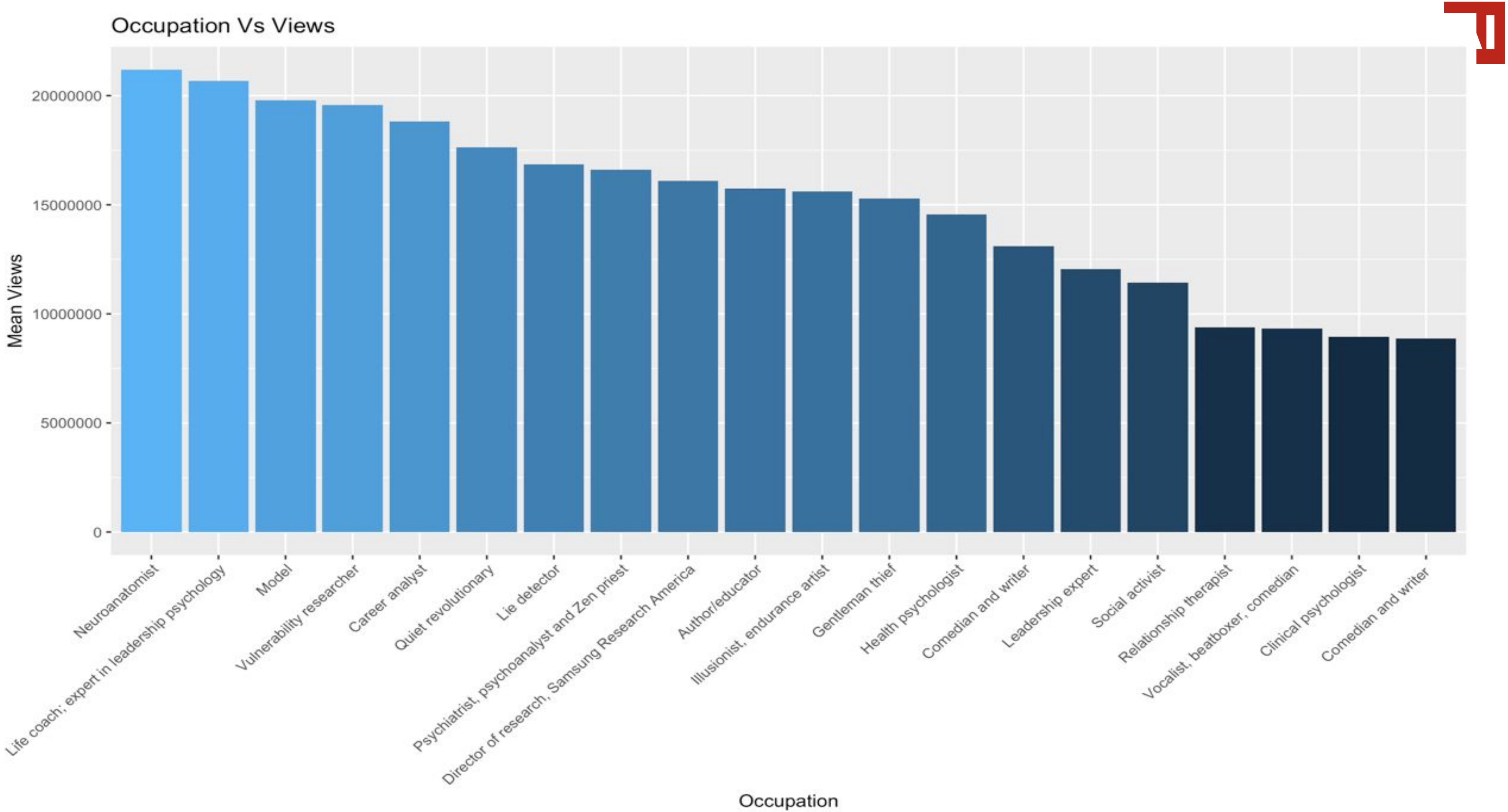


Duration of talks is closer to a normal distribution, but with a wide right-side tail of a few talks at longer duration, around a mean of 14 minutes and median of 12 minutes. Almost all talks range between 1–18 minutes (maximum length of a normal Ted talk).

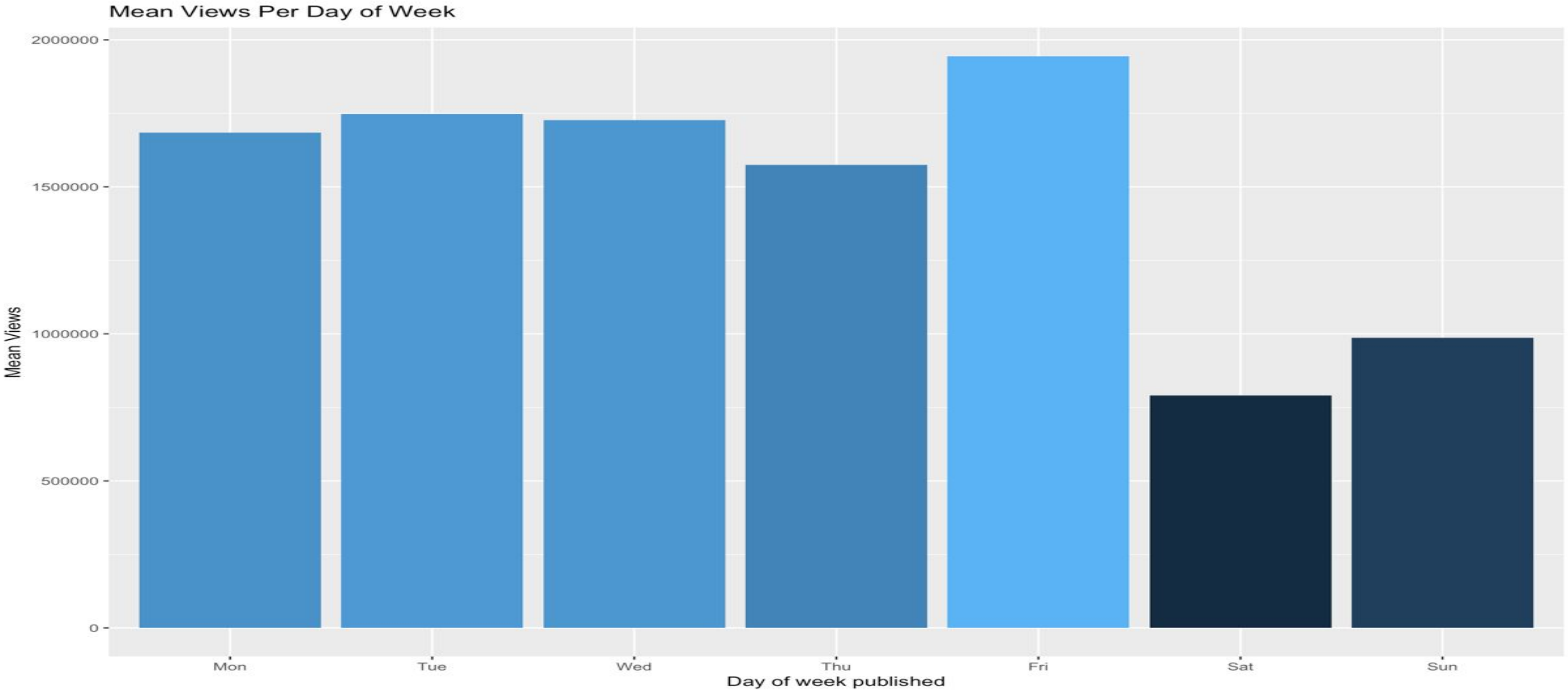


Apparently, writer is the most common occupation for a TED speaker! followed by other creative occupations and

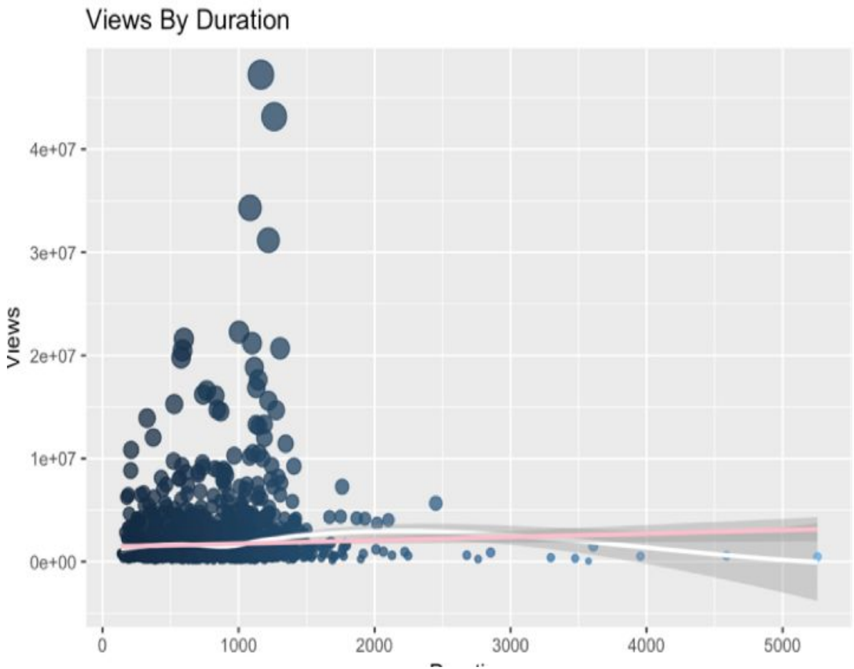
number of speakers with that occupation:

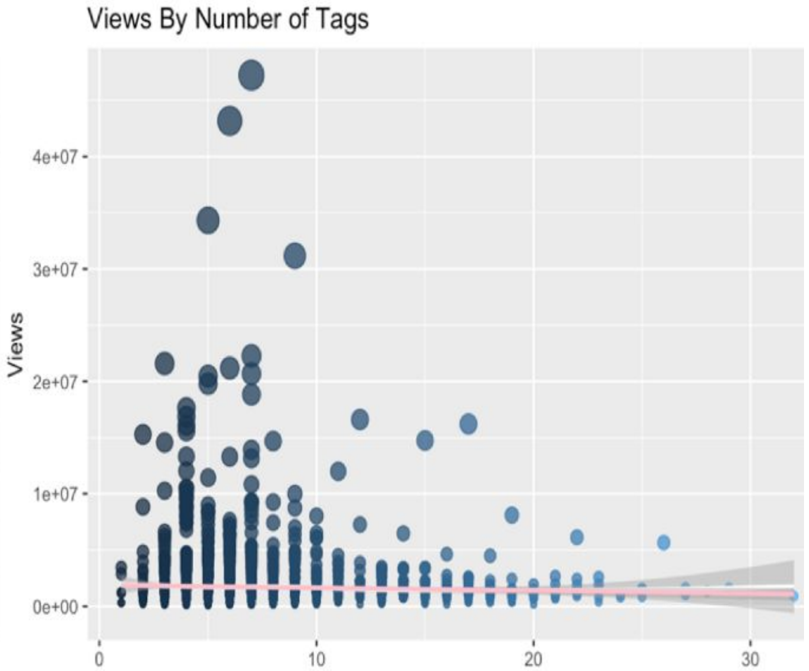


Occupation Vs Views



So, day of week does seem to have some association with average (mean) views! Ted Talks published on weekends seem to have much less views, with Saturday being the lowest, and Friday is the most popular day for ted talks published day.





Unsurprisingly, number of comments is, obviously, very well correlated with number of views, and so does number of languages — they all come from having many viewers. Thus, it is not “fair” to predict views based on these factors, and in the real world, we couldn’t use these parameters to predict, since they **are not causes** for more views, but they are also a result of many views, and a cause in a reinforcing feedback loop: the more comments, the more engaged the community is around the talk and likelier to spread; the more languages, the more viewers can watch; and the more viewers, the more audience there is to comment and translate. The rest had a small linear effect, where that didn’t deviate much, though small.

Conclusion

This limited model conveys correlation, not causation. It does not convey causal relationship well because the fundamental problem of causal inference is not well addressed with these variables, these predictors are *not independent from the y variable,* and they are highly correlated (mostly comments and number of languages which are the best predictors, naturally. I don’t believe that with these available numerical predictors we could have reached a causal inference. Next attempts might use the transcription of the talk to analyze the content, or audio to analyze the level of clapping, or the visuals in the talk and the clothing of the speaker to better predict using the content of the talk.