Project Title: Trend Analysis in TED Talks

Group Number: 5

Submission Date: 12/22/2023

Group Members: Nikhil Gudugula, Siddharth Maredu, Sathish Komire

Signatures of Group Members: Nikhil Gudugula, Siddharth Maredu, Sathish Komire

Table of Contents

Executive Summary

- **Purpose**: The purpose of this project is to analyze key trends in TED Talks, focusing on aspects like the most common words, publication frequency over time, talk durations, speaker occupation types, weekday distribution, and the interaction between views and comments. It showcases proficiency in data analysis and visualization, highlighting how content engagement can be understood in a business context.
- **Content**: Summarize the key aspects of each section of the report, highlighting major findings and conclusions.
- **Tone**: Keep it professional and succinct. This summary should provide a clear snapshot of the project for someone who may not read the entire report.

Introduction

- **Scope**: Our project is explicitly designed to analyze the Trends that happened in Ted Talks from 2006 - 2017. So, to perform this we have selected the Ted talks dataset which contains main information such as the key details of a particular Ted Talk like title, speaker, content, published date, languages, comments, ratings, duration, transcripts etc. To achieve this first we cleaned the datasets and merged the relevant columns that were required for this analysis and we have we have performed our analyses in an iterative way such as like what were the most frequently occurring words for ted talks, the number of ted talks that were published over time, the distribution of ted talks duration, The top shares of talks by speaker occupation type, The number of ted talks published on a weekday basis, the relationship between views and comments for Ted Talks and Total count per Individual Rating Type. Upon this we have created their respective visualizations to demonstrate the trends in Ted Talks.
- **Objectives**: Our goal is to attract and advise investors to join the TED partnership program which allows multiple investors to sponsor a particular TED Talks that aligns with the values of their organization. These investors not only bring money, but they will also be part of the TED Talks culture. To achieve this, we have crafted a series of analyses that would help an investor to choose wisely and appropriately what guidelines they must adhere to so that they can have more outreach under this global platform. They will be having a lasting relationship with TED Talks as their power helps the motive of the TED Talks more relevant and have representation in the real world.

Dataset Selection

Dataset Choice

- **Dataset Name**: TED Talks (Kaggle Link)
- **Reason for Selection**: We wanted to go with this dataset because we were interested in word analysis and what kind of words were mostly spoken in all TED talks combined. Moreover, we have not performed Text Analysis before, so gave us an opportunity to learn this aspect.
- **Confirmation of Dataset Claim**:

This is a snapshot from the Discussion forum

Duplication Penalty Acknowledgment

- **Confirmation of Unique Choice**: We have cross checked all the messages in the discussion forum to avoid duplication of the dataset selection and we hope this serves as proof of dataset selection.

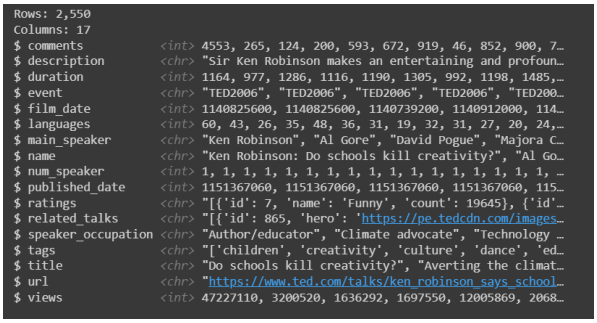
Data Management

Process Detailing

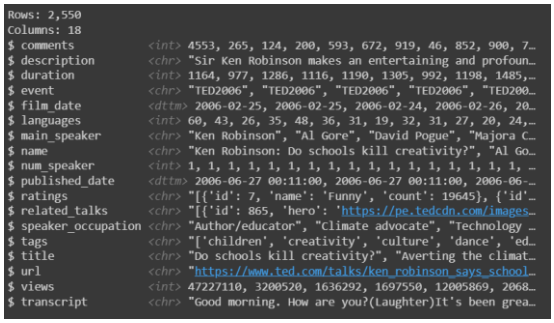
- **Data Cleaning**: Our examination of the dataset focused on its accuracy, consistency, and overall completeness. We encountered some instances of missing data. Given that these missing values constituted a relatively minor portion of the dataset, they were omitted to prevent any skewed analytical results. Also, for the tags column we have removed square brackets and single quotes to make our analyses process easier.
- **Data Preparation**: In the data preparation stage, we merged 'ted_main' and 'ted_transcripts' using a shared 'url' column as this is the unique identifier among these two tables. We also transformed complex structured columns which had multiple tags by splitting them into multiple simpler columns. We have also converted date formats from integers to standard date representations for easier analysis.

Transformation and Integration

- **Data Transformation Techniques**: We implemented transformations on two columns called ratings and related_talks by normalizing nested structures in them into multiple columns for detailed analysis and converting date formats from integers to standard date formats. This facilitated more nuanced analysis.



Before Cleaning



After Cleaning and Merging

- **Data Integration Techniques**: We merged the 'ted_main' and 'ted_transcripts' files using a common 'url' column, employing a left join function from the dplyr package in R. This merge, done in Google Colab, helped us to integrate comprehensive data for holistic analysis.
- **Rationale for Techniques Used**: The chosen techniques enhanced data usability and analysis precision. Unnesting structured columns allowed for granular analysis of each aspect. Merging files using a left join ensured a comprehensive dataset which avoids missing out any rows from the main dataset and thus providing a full picture for our analysis. The transformation of date formats made temporal analysis more intuitive and accurate.

Data Exploration and Analysis

Description and Trends

- **Statistical Descriptions**: Here, we summarized key data points. For instance, the average duration of TED Talks was calculated, revealing how long typical presentations last. We also looked at the distribution of views and comments, noting the average views a TED Talk receives and the typical number of comments. These statistics helped us understand the average engagement level of TED Talks. Additionally, we examined the correlation between different numerical variables, like duration and views, which provided insights into how these factors might influence each other.
- **Identified Trends/Patterns**: In our analysis, we observed a steady rise in TED Talks from 2006 to 2013, correlating with increased digital adoption and internet access. Post-2013, a fluctuation in the number of talks suggested a strategic pivot towards quality over quantity. Our sentiment analysis showed a predominance of positive language across talks. Additionally, the data revealed that while TED Talks generally adhere to a sweet spot in duration, viewership does not strongly predict the number of comments, indicating that factors beyond views drive audience engagement.
- **Contextual Implications**: The findings of our analysis carry significant weight in the context of TED Talks' evolution. The initial surge in talks aligns with the digital boom, indicating that TED leveraged online platforms effectively. The shift towards quality suggests a maturing content strategy. The sentiment analysis underpins TED's positive and inspirational brand image. Lastly, the weak correlation between views and comments signals that engagement is complex, influenced by factors that may include content relevance and the power of ideas presented. These factors will enhance the likelihood of a sponsor to invest in a TED Talk appropriately using the mentioned guidelines, so that a talk can have a good chance of getting good viewership.

Text Analysis

Text Preprocessing

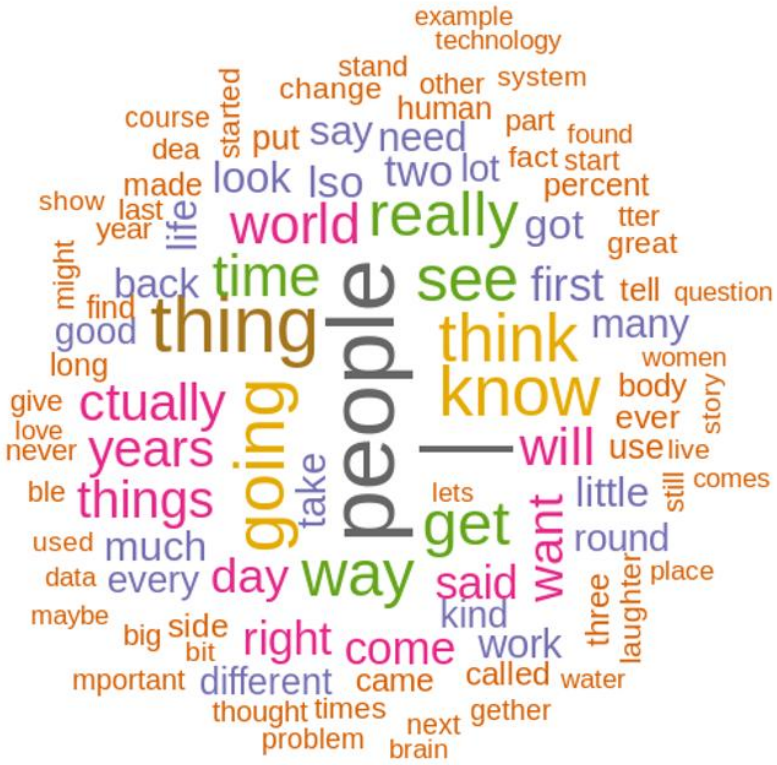
- **Techniques Used**: We applied tokenization to break down text into units for analysis and removed stop words to focus on more meaningful words.

Sentiment Analysis

- **Methodology**: Using R, we calculated sentiment scores by counting the occurrence of positive and negative words.
- **Findings**: The sentiment analysis revealed a higher count of positive words compared to negative, suggesting a generally positive tone in the TED Talks.

Topic Modeling

- **Approach**: Used methods like word frequency analysis to identify common themes.
- **Results**: The word cloud indicates prevalent topics focused on people, technology, and personal experiences.



Word Cloud

Interpretation

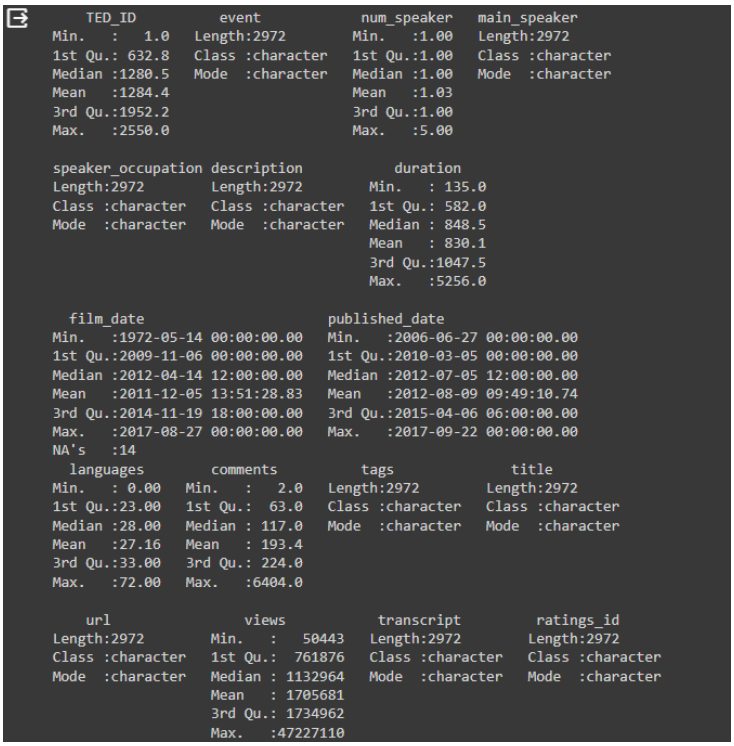
- **Implications of Text Analysis Results**: Implications of Text Analysis Results: The positive sentiment and the identified topics reflect TED Talks' focus on inspiring and innovative content.

Statistical Modeling

Model Application

- **Model Used**: We utilized a linear regression model to analyze the relationship between views and comments on TED Talks.

- **Assumptions Made**: We assumed a linear relationship between the variables, homoscedasticity, and that the residuals are normally distributed.
- **Model Results**: The model indicated a positive correlation between views and comments, although the correlation coefficient was relatively low.



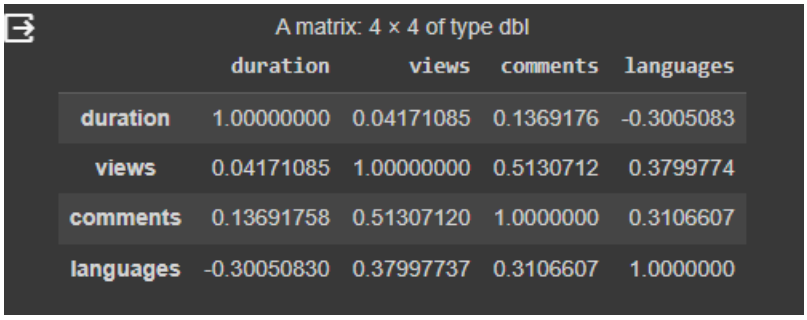
```
     TED_ID          event           num_speaker   main_speaker
 Min.   :   1.0   Length:2972        Min.   :1.00   Length:2972
 1st Qu.: 632.8   Class :character   1st Qu.:1.00   Class :character
 Median :1280.5   Mode  :character   Median :1.00   Mode  :character
 Mean   :1284.4                     Mean   :1.03
 3rd Qu.:1952.2                     3rd Qu.:1.00
 Max.   :2550.0                     Max.   :5.00

 speaker_occupation description           duration
 Length:2972        Length:2972        Min.   : 135.0
 Class :character   Class :character   1st Qu.: 582.0
 Mode  :character   Mode  :character   Median : 848.5
                                       Mean   : 830.1
                                       3rd Qu.:1047.5
                                       Max.   :5256.0

   film_date                   published_date
 Min.   :1972-05-14 00:00:00.00   Min.   :2006-06-27 00:00:00.00
 1st Qu.:2009-11-06 00:00:00.00   1st Qu.:2010-03-05 00:00:00.00
 Median :2012-04-14 12:00:00.00   Median :2012-07-05 12:00:00.00
 Mean   :2011-12-05 13:51:28.83   Mean   :2012-08-09 09:49:10.74
 3rd Qu.:2014-11-19 18:00:00.00   3rd Qu.:2015-04-06 06:00:00.00
 Max.   :2017-08-27 00:00:00.00   Max.   :2017-09-22 00:00:00.00
 NA's   :14
   languages        comments          tags              title
 Min.   : 0.00   Min.   :   2.0   Length:2972        Length:2972
 1st Qu.:23.00   1st Qu.:  63.0   Class :character   Class :character
 Median :28.00   Median : 117.0   Mode  :character   Mode  :character
 Mean   :27.16   Mean   : 193.4
 3rd Qu.:33.00   3rd Qu.: 224.0
 Max.   :72.00   Max.   :6404.0

     url                views           transcript         ratings_id
 Length:2972        Min.   :   50443   Length:2972        Length:2972
 Class :character   1st Qu.:  761876   Class :character   Class :character
 Mode  :character   Median : 1132964   Mode  :character   Mode  :character
                    Mean   : 1705681
                    3rd Qu.: 1734962
                    Max.   :47227110
```

Dataset Summary



A matrix: 4 × 4 of type dbl

	duration	views	comments	languages
duration	1.00000000	0.04171085	0.1369176	-0.3005083
views	0.04171085	1.00000000	0.5130712	0.3799774
comments	0.13691758	0.51307120	1.0000000	0.3106607
languages	-0.30050830	0.37997737	0.3106607	1.0000000

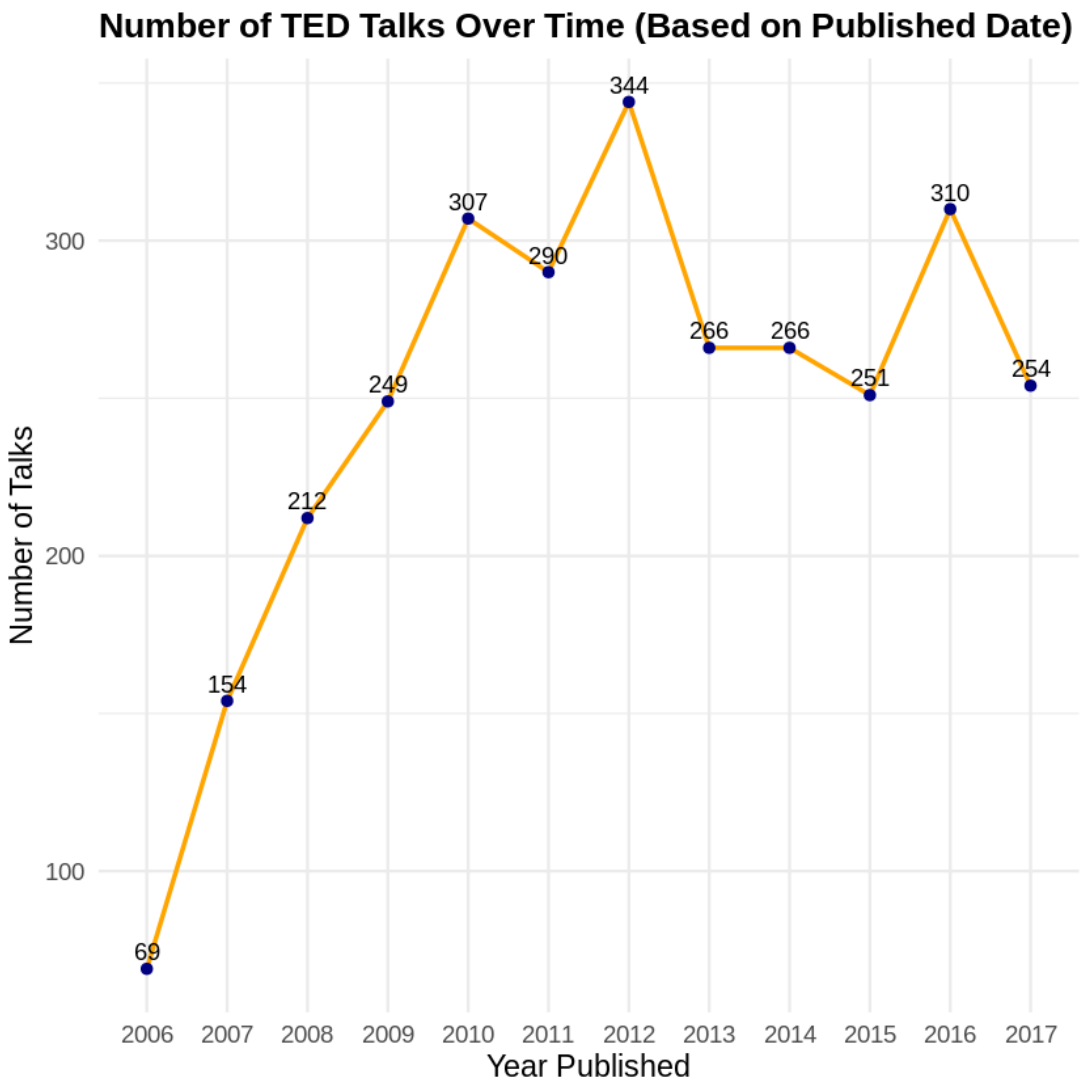
Correlation Matrix

- **Implications of Model Results**: The results suggest that while there is a positive relationship between the number of views and comments, it's not strongly predictive. This implies that factors other than views may influence the number of comments a TED Talk receives.

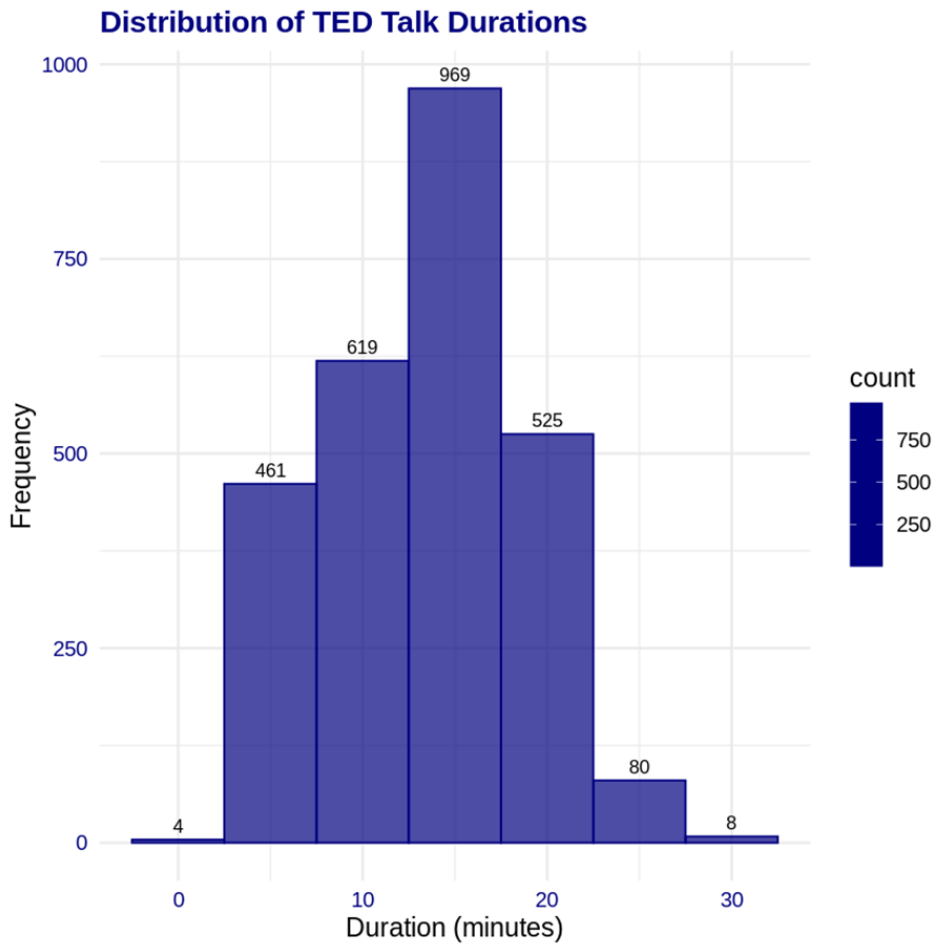
Data Visualization

Visualization Development

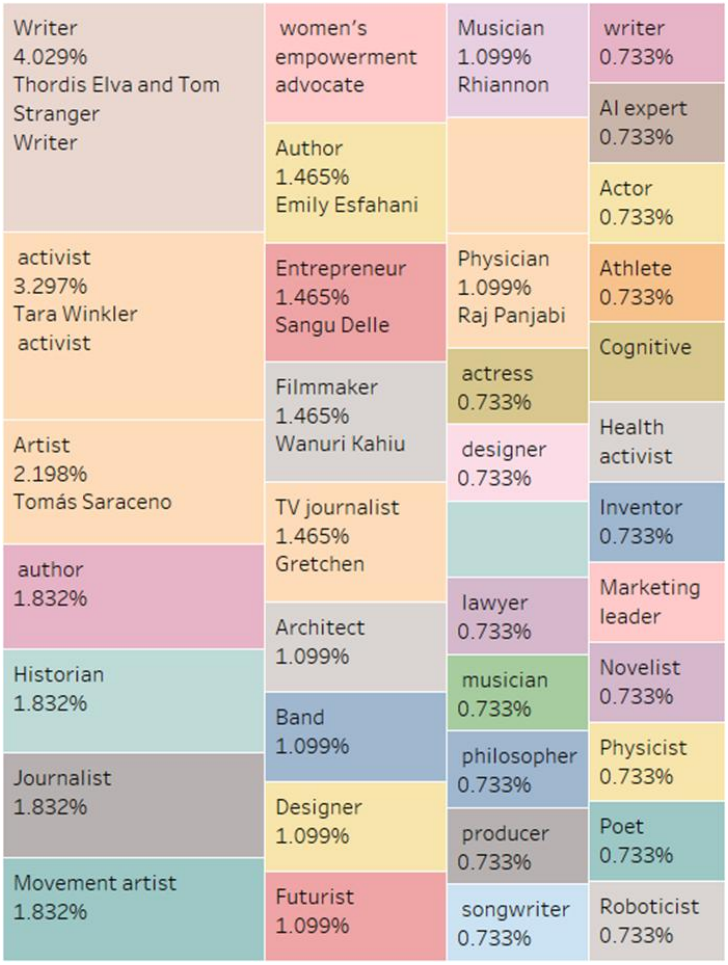
- **Tools Used**: We have used Power Query for cleaning and filtering, R for data wrangling and analysis and Tableau for visualization
- **Description of Visualizations**:



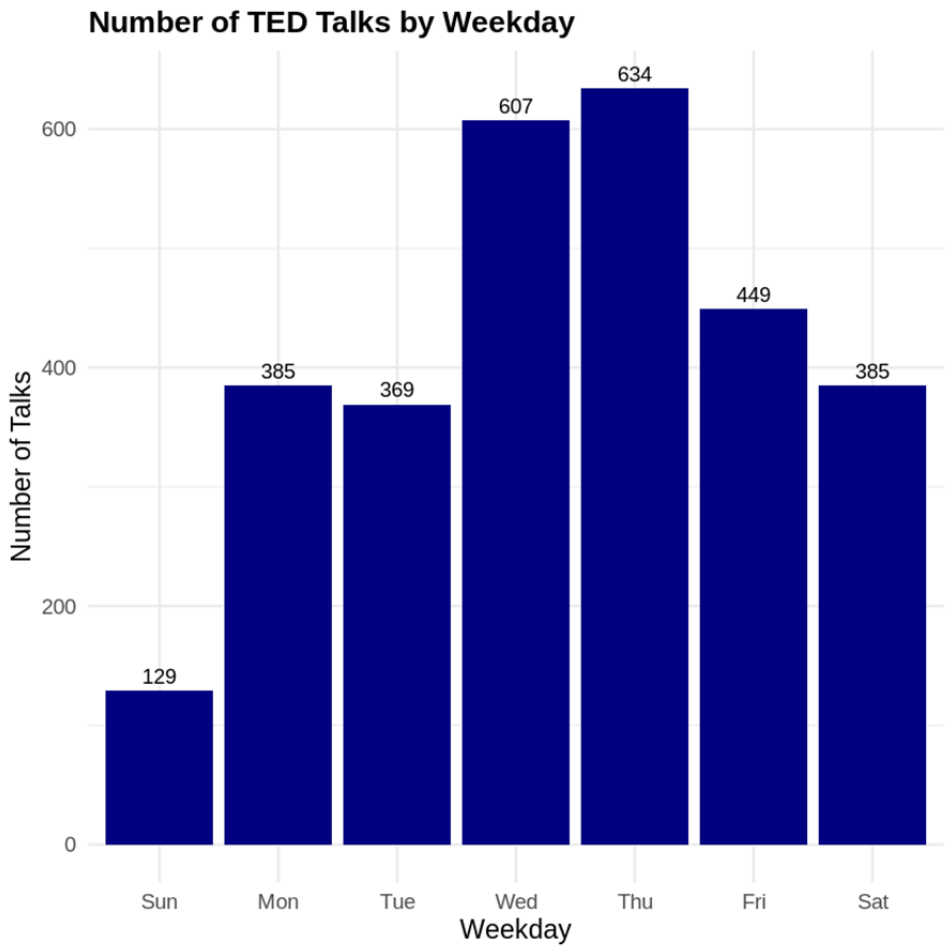
- **Interpretation**: The graph indicates a consistent rise in the production of TED Talks from 2006 to 2013, aligning with increased digitalization and accessibility to the internet. Post-2013, there's a noticeable variability in the number of talks, reflecting TED's shift in strategy towards prioritizing the quality of content over the sheer volume of talks produced.



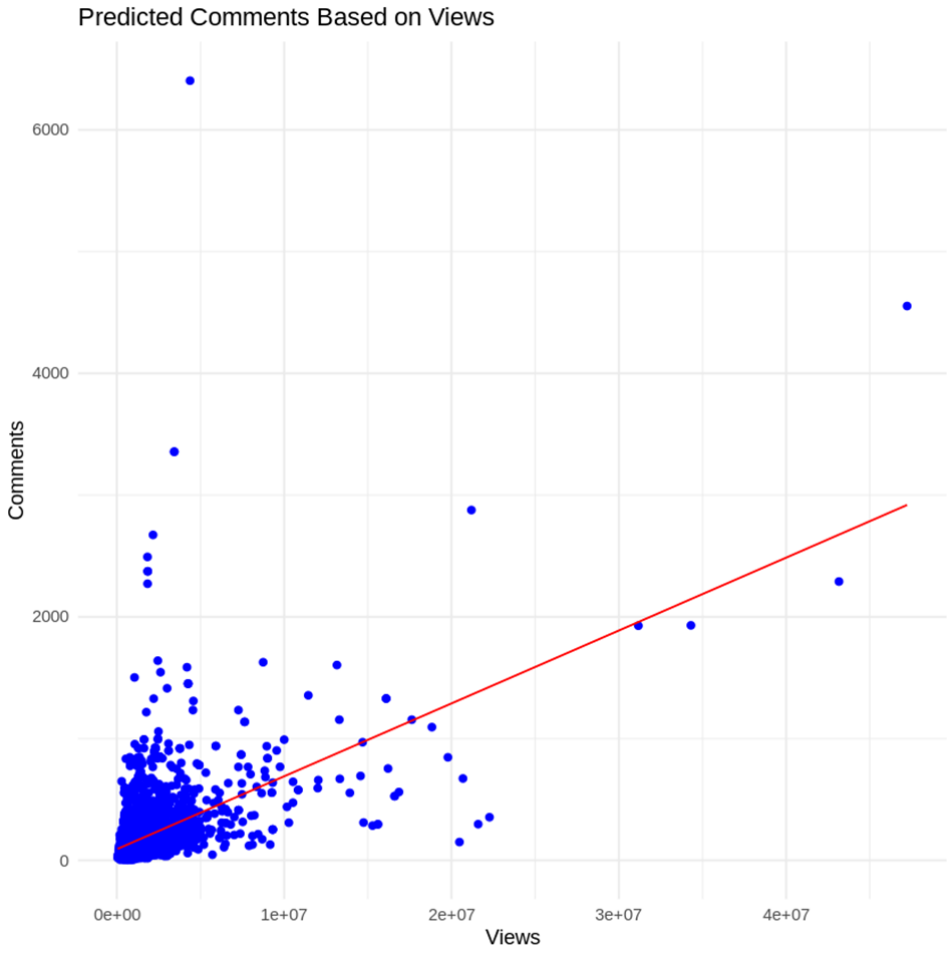
- **Interpretation**: The histogram displays the distribution of TED Talk durations. Most talks last around 10 to 20 minutes, with the highest frequency around 10 minutes, indicating a preference for talks within this duration. There are notably fewer talks that are either very short (under 5 minutes) or longer (over 20 minutes), suggesting that the TED format favors concise, impactful presentations.



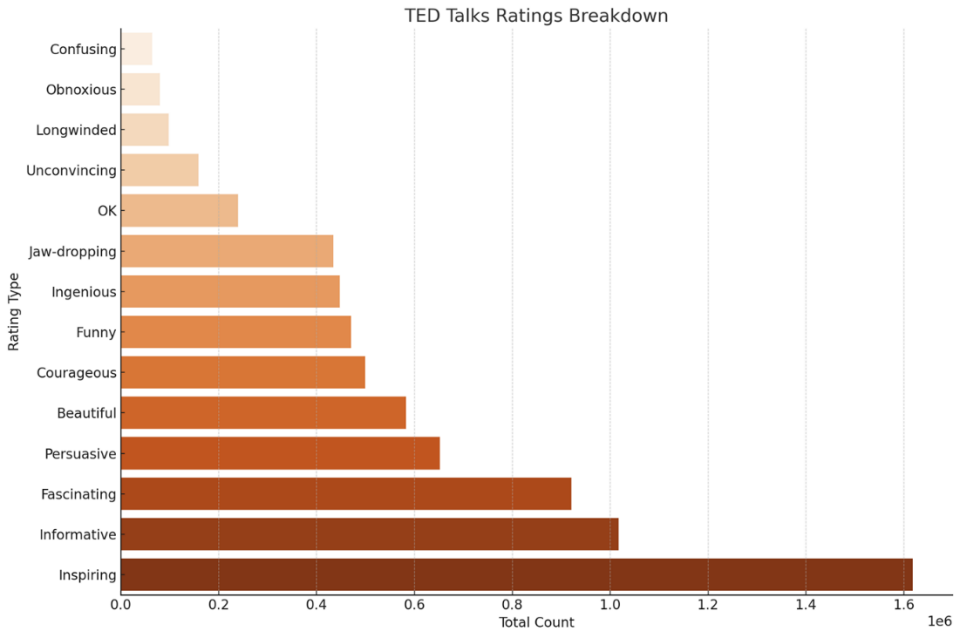
- **Interpretation**: As we can see, the tree map visualization representing the distribution of speaker occupations in TED Talks. Each block's size and label indicate the proportion of talks given by speakers from various professions, with writers and activists representing the larger shares. This suggests a diverse range of perspectives featured at TED events, emphasizing the platform's interdisciplinary nature. We can also see the popular Artist's name for each category



- **Interpretation**: The chart reveals a week in the life of TED Talks, almost telling a story. It starts quietly on Sunday, a day of rest with the fewest talks. As the new week rolls in, we see a climb in numbers, peaking midweek when minds are active, and audiences are perhaps seeking inspiration.



- **Interpretation**: The plot depicts the relationship between the views and comments of TED Talks. It shows a scattered yet positive trend where, generally, as views increase, comments tend to increase too. However, the relationship is not strong; many talks with a high number of views don't necessarily have a high number of comments. This suggests that while views are an indicator of popularity, they don't always guarantee audience interaction in the form of comments



- **Interpretation**: The bar chart "TED Talks Ratings Breakdown" displays various ratings given to TED Talks, ordered from the least frequent at the top to the most frequent at the bottom. The most common rating is "Inspiring", followed by "Informative" and "Fascinating", indicating that viewers are most impacted by talks that motivate, educate, and interest them. Less common ratings are negative, such as "Confusing" and "Obnoxious", suggesting that such feedback is relatively rare. This distribution reflects the positive reception and educational nature of TED Talks.

- **Adherence to Best Practices**:
  Our visualizations were crafted following best practices for data presentation, ensuring each graph and chart is clear, precise, and easy to interpret. Color schemes were chosen for optimal contrast and accessibility, labels and legends were made explicit for immediate understanding, and data was presented without distortion to honestly reflect the true nature of the findings. Each visualization serves a purpose, helping viewers quickly grasp the story behind the numbers.

Visualization Interpretation

- **Meaning of Visualizations**: We have given the interpretation just right below each visualization, so we think that we can omit this section.

- **Relevance to Objectives**: The visualizations in our project directly support our objectives by providing potential investors with a clear understanding of TED Talks' dynamics. The growth trend over time, sentiment analysis, and engagement patterns offer insights into the platform's reach and influence. This data allows investors to identify successful content themes, speaker profiles, and engagement strategies, aligning their sponsorship with talks that reflect their values

and have the potential for significant impact, fostering a symbiotic relationship with TED's mission and culture.

Conclusion

- **Summary of Findings**: Our analyses revealed a peak in TED Talks production aligned with digital growth, a preference for content reflecting positive themes, and a nuanced relationship between viewership and audience interaction.
- **Significance**: These insights are crucial for investors looking to align with TED's influential platform, ensuring their contributions support content that resonates with global audiences and embodies their organizational values.
- **Limitations/Recommendations**: The scope of data may have temporal limitations; future research could include updated trends and a broader range of engagement metrics to refine investment strategies further.

References

- **Citing Sources**: List all the references in a consistent format (APA/MLA as required by your course). Include all sources used for data, methods, and theoretical frameworks.

Appendix

- Include your presentation slides and detailed speaker notes.

Submission Checklist (check off each box completed)

- ☐ Project title and group details are complete.
- ☐ All sections are thoroughly addressed.
- ☐ Data visualizations are included and properly referenced.
- ☐ The report adheres to the specified format and guidelines.
- ☐ The report is free from grammatical errors and typos.
- ☐ An appendix with the slides and detailed speaker notes is included.

All members contributed to t