

# **Data Science and Artificial Intelligence in Industrial Work I**

TED Talks Viewership Analysis- A CRISP-DM Approach

---

Tanishq Quraishi

January 15, 2025

## About the Business

---

## **Squirrel Studios**

A marketing and branding agency focused on data-driven strategies.

## About the Client

---

## **TED: Technology, Entertainment, and Design**


- A non-profit organization focusing on Technology, Entertainment, and Design.
- Mission: Spread powerful ideas to spark conversation and drive change.
- Key Metrics:
  - 2B global views
  - 180M monthly followers.

## **Ted Talks: Discover ideas worth spreading**

- Talks on Technology, Entertainment and Design – plus science, business, global issues, the arts and more.

# The Brand

**TED** Ideas change everything

WATCH DISCOVER ATTEND PARTICIPATE ABOUT SIGN IN **MEMBERSHIP** 

Q Search for a talk Sort By Topics Subtitles Duration


## TED Talks: Discover ideas worth spreading

TED-Ed Psychology Leadership Education AI Sleep Mental Health


Business Motivation Communication Personal Growth Sports Health Language

[See all](#)


### Newest Talks




**Solar energy is even cheaper than you think**  
JENNY CHASE 08:59



**Does your heartbeat shape your sense of time?**  
IRINA ARSLANOVA 09:25



**The potential US TikTok ban — and what's at stake**  
CLAY SHIRKY 1:00:52



**Can AI companions help heal loneliness?**  
EUGENIA KUYDA 11:44

Source: <https://www.ted.com/talks>

## Client Query

---

## Objective of the Analysis:

- TED wants to investigate whether the phenomenon of shorter attention spans in humans has reduced the viewership of their videos.

## Specific Focus Areas:

- The potential impact of short-form media platforms such as:
  - **TikTok:** Launched in 2016.
  - **Instagram Reels:** Launched in 2020.



# Project Goals

---

# Project Goals

## **Business Goal**

Help TED improve viewership of TED Talks videos and establish Squirrel Studios as a leader in brand data analytics.

## **Project Goal**

Explore how factors such as duration influence TED Talks viewer engagement.

# Project Stages

---

# Project Stages

Project stages involve:

1. Centralize the client's goal → Data analytics to assess the relationship between duration and viewership.
2. Request for resources from the client → TED Talks dataset.
3. Process and analyze resources → Tech stack: Python, Pandas, Numpy, VS Code, Git.
4. Gather actionable insights through internal evaluations.
5. Submit a report to the client → Including suggestions made by Squirrel Studios.

## About the Data

---

# About the Data

## Dataset Overview:

- Contains data on all TED Talks until April 18th, 2020.
- Main file: `ted_main.csv`.
- Features include views, tags, posted date, speakers, and titles.

**Acknowledgements:** Data scraped from the official TED website.

<sup>1</sup>

**License:** Copyright © AFatani. All rights reserved. TED's videos may be used for non-commercial purposes under a Creative Commons License

---

<sup>1</sup><https://www.kaggle.com/datasets/ahmadfatani/ted-talks-dataset?resource=download>

# Data Understanding

---

# Data Understanding (Table 1)

**Table 1:** Data Categories

Feature	Description
title	Title of the TED Talk.
speaker_name	Name of the speaker who delivered the talk.
views	Total number of views the talk received (engagement metric).
tags	Comma-separated list of tags describing themes or topics of the talk.
duration	Length of the talk in MM:SS format (convertible to seconds).
posted_date	Date the talk was published, in MMM YYYY format.
about_talk	Brief description of the talk content.
about_speaker	Information about the speaker, such as their background or expertise.



**Table 2:** Category Types

<b>Feature Type</b>	<b>Features</b>
Text-Based Features	title, tags, about_talk, about_speaker
Numerical Features	views, duration
Date Feature	posted_date

## Top 5 Most Viewed Videos

**Table 3:** Top 5 Most Viewed Videos

Title	Views
This is what happens when you reply to spam email	60,237,459
Inside the mind of a master procrastinator	40,135,933
The next outbreak? We're not ready	36,342,453
My philosophy for a happy life	35,114,993
What makes a good life? Lessons from the longe...	34,095,862

## Top 5 Least Viewed Videos

**Table 4:** Top 5 Least Viewed Videos

<b>Title</b>	<b>Views</b>
Cómo usar el arte de la fotografía para restau...	10,231
"Illusions for a better society"	10,687
Undocumented lives, inside out	10,793
Humanity at the intersection of science and ar...	11,028
Por qué necesitamos proteger el alta mar	11,411

## Data Cleaning Steps:

- **Cleaning the views Column:**

- Replaced invalid entries such as empty strings, "N/A," or "None" with NaN.
- Removed commas from numerical entries (e.g., 60,237,459).
- Converted the column to a numeric format, with invalid entries coerced to NaN.

- **Cleaning the duration Column:**

- Created a function to convert MM:SS format into seconds.
- Invalid formats were handled by assigning NaN.

## Metrics Before and After Cleaning:

**Table 5:** Metrics Before and After Cleaning

Metric	Before Cleaning	After Cleaning
Number of Unique Videos	2160	2155
Number of Unique Speakers	1915	1912

# Average Views and Duration

## Key Statistics:

- **Average Number of Views:** 2,060,890.82
- **Average Duration:** 658.50 seconds (approximately 11 minutes)

# Modeling and Evaluation

---

# Modeling and Evaluation - Part 1

**Objective:** Assess the influence of talk duration on TED Talks viewership using statistical modeling.

## Steps Taken:

### 1. Data Normalization:

- Normalized duration and views to improve model convergence by scaling it

### 2. Model Building:

- Used a Mixed Linear Model (LMM) to predict views based on `duration_scaled`.
- Included `speaker_name` as a random effect to account for speaker-specific effects.

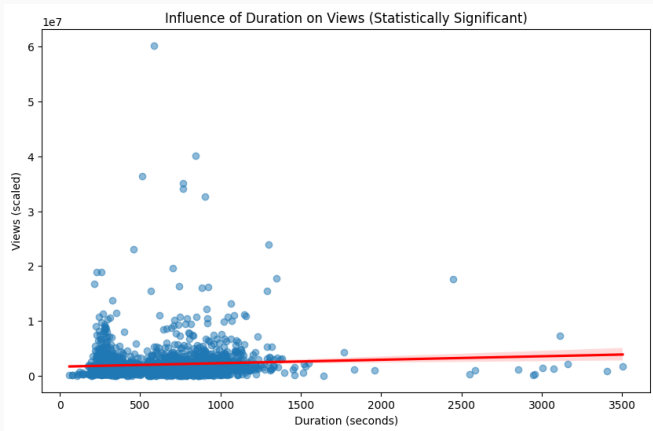


### Steps Taken (continued):

#### 3. Model Evaluation:

- Examined statistical significance of duration:
  - **p-value:** 0.004 (statistically significant at 0.05 level).

## Key Insights:



## Key Insights:

- The p-value for duration (0.004) is below the common significance threshold of 0.05.
- This indicates that the relationship between video duration (scaled) and normalized views is statistically significant.
- Duration of a talk positively influences its viewership.

# **Project Report and Suggestions**

---

## Recommendations on TED Talk Viewership:

- **Content Optimization:**

- Ensure talks are engaging and long, aligning with the duration analysis.

## Areas for Further Analysis:

- **Impact of Translations:** Investigate how translated titles and subtitles affect viewership and identify best practices for localization.
- **Engagement Metrics:** Analyze other engagement metrics, such as likes, shares, and comments, to gain a comprehensive understanding of audience interaction.