
TED
TALKS

1 DATA MANIPULATION USING PANDAS

1.1 INTRODUCTION

This document serves as a comprehensive guide for working with the TED Talks dataset using the pandas library in Python. By leveraging the powerful functionalities of pandas, we will explore the dataset, derive new columns, handle missing values, and export the modified data. pandas enables efficient data manipulation, analysis, and provides valuable insights for making informed decisions based on the TED Talks dataset or any other tabular data. Throughout this document, we will delve into various operations and methods, providing step-by-step explanations and examples to facilitate a comprehensive understanding of working with the TED Talks dataset using pandas.

1.2 IMPORTING FILE AND READING THE FILE

`import pandas as pd` : This line imports the pandas library, which is a powerful data manipulation and analysis tool in Python. It is commonly aliased as `pd` for convenience.

`df = pd.read_csv('TED Talks.csv')` : This line reads the CSV file named 'TED Talks.csv' using the `read_csv()` function from pandas. The file should be in the same directory as your Python script or Jupyter Notebook. The resulting data is stored in a pandas DataFrame named `df` , which is a tabular data structure that allows you to perform various operations and analysis.

In [1]:

```
import pandas as pd
```

In [2]:

```
df=pd.read_csv('TED Talks.csv')
```

1.3 VIEWING THE FIRST 5 ROWS OF THE DATASET

The df.head() command returns the first 5 rows of the DataFrame df. It is a convenient way to quickly examine the structure and contents of the DataFrame. By default, it displays the top 5 rows, but you can specify a different number by passing the desired value as an argument to head()

In [3]:

```
df.head()
```

Out[3]:

	title	author	date	views	likes	
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	December 2021	404000.0	12000.0	https://ted.com/talks/ozawa_bineshi_a
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	February 2022	214000.0	6400.0	https://ted.com/talks/sydney_iaukea_the
2	How play can spark new ideas for your business	Martin Reeves	September 2021	412000.0	12000.0	https://ted.com/talks/martin_reeves_hov
3	Why is China appointing judges to combat clima...	James K. Thornton	October 2021	427000.0	12000.0	https://ted.com/talks/james_k_thornton_
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	October 2021	2400.0	72.0	https://ted.com/talks/mahendra_singhi_c

1.4 SHAPE AND INFO ABOUT THE DATASET

`df.shape` returns a tuple containing the dimensions of the DataFrame, representing the number of rows and columns respectively.

The `df.info()` command displays a summary of the DataFrame's structure, including the number of non-null values in each column, the data type of each column, and the total memory usage. This method is commonly used to gain an overview of the dataset's contents and assess missing values.

In [4]:

```
df.shape
```

Out[4]:

```
(5444, 6)
```

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5444 entries, 0 to 5443
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   title   5444 non-null   object 
 1   author  5439 non-null   object 
 2   date    5440 non-null   object 
 3   views   5440 non-null   float64
 4   likes   5440 non-null   float64
 5   link    5440 non-null   object 
dtypes: float64(2), object(4)
memory usage: 255.3+ KB
```

1.5 CHECKING FOR NULL VALUES IN DATASET

The `df.isnull().sum()` method in pandas is used to calculate the total number of missing or null values in each column of a DataFrame.

In [6]:

```
df.isnull()
```

Out[6]:

	title	author	date	views	likes	link
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
5439	False	False	False	False	False	False
5440	False	True	True	True	True	True
5441	False	True	True	True	True	True
5442	False	True	True	True	True	True
5443	False	True	True	True	True	True

5444 rows × 6 columns

The `df.isnull().sum()` command calculates the number of missing or null values in each column of the DataFrame `df`. It returns a Series where each column name is paired with the respective count of null values.

In [7]:

```
df.isnull().sum()
```

Out[7]:

```
title     0
author    5
date      4
views     4
likes     4
link      4
dtype: int64
```

The `df.isnull().sum(axis=1)` method in pandas is used to calculate the total number of missing or null values in each row of a DataFrame.

In [8]:

```
df.isnull().sum(axis=1)
```

Out[8]:

```
0       0
1       0
2       0
3       0
4       0
       ..
5439    0
5440    5
5441    5
5442    5
5443    5
Length: 5444, dtype: int64
```

These methods are commonly used to identify missing data and assess data quality in a DataFrame. By analyzing null values, it becomes easier to make decisions about handling missing data, such as imputation or removing incomplete rows/columns.

1.6 DROPPING NULL VALUES

The `df.dropna()` method in pandas is used to remove missing or null values from a DataFrame. It returns a new DataFrame with rows or columns containing null values removed.

This method is commonly used to clean the data by removing incomplete or missing information. By dropping rows or columns with null values, it helps to ensure data quality and facilitates subsequent analysis.

In [9]:

```
df=df.dropna()
```

In [12]:

```
df.isnull().sum()
```

Out[12]:

```
title     0
author    0
date      0
views     0
likes     0
link      0
dtype: int64
```

In [13]:

```
df
```

Out[13]:

	title	author	date	views	likes	
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	December 2021	404000.0	12000.0	https://ted.com/talks/ozawa_
1	The dark history of the overthrow of Hawaii	Sydney laukea	February 2022	214000.0	6400.0	https://ted.com/talks/sydney_ia
2	How play can spark new ideas for your business	Martin Reeves	September 2021	412000.0	12000.0	https://ted.com/talks/martin_re
3	Why is China appointing judges to combat clima...	James K. Thornton	October 2021	427000.0	12000.0	https://ted.com/talks/james_k_
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	October 2021	2400.0	72.0	https://ted.com/talks/mahendra_
...	...	...	...	...	...	
5435	The best stats you've ever seen	Hans Rosling	February 2006	15000000.0	458000.0	https://ted.com/talks/hans_ros
5436	Do schools kill creativity?	Sir Ken Robinson	February 2006	72000000.0	2100000.0	https://ted.com/talks/sir_ken_rc
5437	Greening the ghetto	Majora Carter	February 2006	2900000.0	88000.0	https://ted.com/talks/majora_
5438	Simplicity sells	David Pogue	February 2006	2000000.0	60000.0	https://ted.com/talks/david_pc
5439	Averting the climate crisis	Al Gore	February 2006	3600000.0	109000.0	https://ted.com/talks/al_gore

5439 rows × 6 columns

1.7 CHECKING FOR DUPLICATE VALUES

The `df.duplicated().any()` command checks if there are any duplicate rows in the DataFrame `df`. It returns a boolean value indicating whether any duplicates are found (`True`) or not (`False`).

This method is useful for identifying and handling duplicate data. By checking for duplicates, you can ensure data integrity and accuracy. Further actions can be taken, such as removing duplicate rows, dropping unnecessary repetitions, or investigating potential data entry errors.

In [14]:

```
df.duplicated().any()
```

Out[14]:

```
False
```

1.8 CHECKING FOR UNIQUE VALUES IN COLUMNS

The `df['column'].nunique()` command calculates the number of unique values in the column specified by `'column'` within the DataFrame `df`. It returns the count of distinct values.

This method is commonly used to analyze the diversity or variability of data within a specific column. By counting the number of unique values, it provides insights into the distinct entities or categories present in the column. For example, `df['title'].nunique()` would return the count of unique titles in the `'title'` column, while `df['author'].nunique()` would return the count of unique authors in the `'author'` column.

In [15]:

```
df['title'].nunique()
```

Out[15]:

```
5439
```

In [16]:

```
df['author'].nunique()
```

Out[16]:

```
4443
```

1.9 CONVERTING OBJECT DATATYPE INTO DATE DATATYPE

The `df.dtypes` command is used to retrieve the data types of each column in the DataFrame `df`. It returns a Series that displays the data type of each column.

The `pd.to_datetime()` function in pandas is used to convert a column or series containing dates or date-like strings into datetime objects. It allows for convenient manipulation and analysis of date and time data.

By using `pd.to_datetime()`, we ensure that the data in the `'date'` column is treated as proper datetime objects, which is useful for time-based analysis and visualization.

In [17]:

```
df.dtypes
```

Out[17]:

```
title      object
author     object
date       object
views     float64
likes     float64
link       object
dtype: object
```

In [20]:

```
df['date']=pd.to_datetime(df['date'])
```

In [21]:

```
df.dtypes
```

Out[21]:

```
title             object
author            object
date      datetime64[ns]
views            float64
likes            float64
link              object
dtype: object
```

1.10 Extracting Year and Month from Date Values and Creating Separate Columns

The `.dt.year` accessor in pandas is used to extract the year component from a datetime column. It allows you to access the year values from a datetime column and store them in a new column.

The `df['date'].dt.year` command extracts the year component from the `'date'` column of the DataFrame `df` using the `.dt.year` accessor. It assigns the extracted year values to a new column named `'year'`.

In [22]:

```
df['year']=df['date'].dt.year
```

The `.dt.month_name()` accessor in pandas is used to extract the month name component from a datetime column. It allows you to access the month names from a datetime column and store them in a new column.

The `df['date'].dt.month_name()` command extracts the month name component from the `'date'` column of the DataFrame `df` using the `.dt.month_name()` accessor. It assigns the extracted month names to a new column named `'month'`.

In [23]:

```
df['month']=df['date'].dt.month_name()
```

1.11 DROPPING DATE COLUMN

The `df.drop('column', axis=1)` method in pandas is used to remove a specified column from a DataFrame. It returns a new DataFrame with the specified column dropped.

The `df.drop('date', axis=1)` command removes the `'date'` column from the DataFrame `df`. It returns a new DataFrame without the dropped column.

In [25]:

```
df=df.drop('date',axis=1)
```

In [26]:

```
df
```

Out[26]:

	title	author	views	likes	
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	404000.0	12000.0	https://ted.com/talks/ozawa_bineshi_albe
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	214000.0	6400.0	https://ted.com/talks/sydney_iaukea_the_da
2	How play can spark new ideas for your business	Martin Reeves	412000.0	12000.0	https://ted.com/talks/martin_reeves_how_pl
3	Why is China appointing judges to combat clima...	James K. Thornton	427000.0	12000.0	https://ted.com/talks/james_k_thornton_why
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2400.0	72.0	https://ted.com/talks/mahendra_singhi_ceme
...	...	...	...	...	
5435	The best stats you've ever seen	Hans Rosling	15000000.0	458000.0	https://ted.com/talks/hans_rosling_the_bes
5436	Do schools kill creativity?	Sir Ken Robinson	72000000.0	2100000.0	https://ted.com/talks/sir_ken_robinson_do_s
5437	Greening the ghetto	Majora Carter	2900000.0	88000.0	https://ted.com/talks/majora_carter_greeni
5438	Simplicity sells	David Pogue	2000000.0	60000.0	https://ted.com/talks/david_pogue_simplic
5439	Averting the climate crisis	Al Gore	3600000.0	109000.0	https://ted.com/talks/al_gore_averting_th

5439 rows × 7 columns

1.12 CREATING NEW COLUMNS USING EXISTING COLUMNS

The code `df['Engagement ratio'] = df['views'] / df['likes']` creates a new column named `'Engagement ratio'` in the DataFrame `df`. The new column is derived by performing a mathematical operation that divides the values in the `'views'` column by the values in the `'likes'` column.

The `df['views'] / df['likes']` calculation divides the values in the `'views'` column by the corresponding values in the `'likes'` column, resulting in the engagement ratio. This ratio represents the number of views per like for each TED Talk.

By creating a new column based on existing columns, we can derive additional meaningful insights from your data. In this case, the engagement ratio column provides a metric to evaluate the level of audience engagement for each TED Talk.

In [27]:

```
df['Engagement ratio']=df['views']/df['likes']
```

In [28]:

```
df
```

Out[28]:

	title	author	views	likes	
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	404000.0	12000.0	https://ted.com/talks/ozawa_bineshi_albe
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	214000.0	6400.0	https://ted.com/talks/sydney_iaukea_the_da
2	How play can spark new ideas for your business	Martin Reeves	412000.0	12000.0	https://ted.com/talks/martin_reeves_how_pl
3	Why is China appointing judges to combat clima...	James K. Thornton	427000.0	12000.0	https://ted.com/talks/james_k_thornton_why
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2400.0	72.0	https://ted.com/talks/mahendra_singhi_ceme
...	...	...	...	...	
5435	The best stats you've ever seen	Hans Rosling	15000000.0	458000.0	https://ted.com/talks/hans_rosling_the_bes
5436	Do schools kill creativity?	Sir Ken Robinson	72000000.0	2100000.0	https://ted.com/talks/sir_ken_robinson_do_s
5437	Greening the ghetto	Majora Carter	2900000.0	88000.0	https://ted.com/talks/majora_carter_greeni
5438	Simplicity sells	David Pogue	2000000.0	60000.0	https://ted.com/talks/david_pogue_simplic
5439	Averting the climate crisis	Al Gore	3600000.0	109000.0	https://ted.com/talks/al_gore_averting_th

5439 rows × 8 columns

1.13 SAVING THE FILE

The `df.to_csv('filename.csv')` method in pandas is used to save a DataFrame to a CSV (Comma-Separated Values) file. It writes the contents of the DataFrame to the specified file path in CSV format.

The `df.to_csv('TED Talks-Viz.csv')` command saves the DataFrame `df` to a CSV file named `'TED Talks-Viz.csv'`. The resulting file will be created in the same directory as the notebook.

In [29]:

```
df.to_csv('TED Talks-Viz.csv')
```

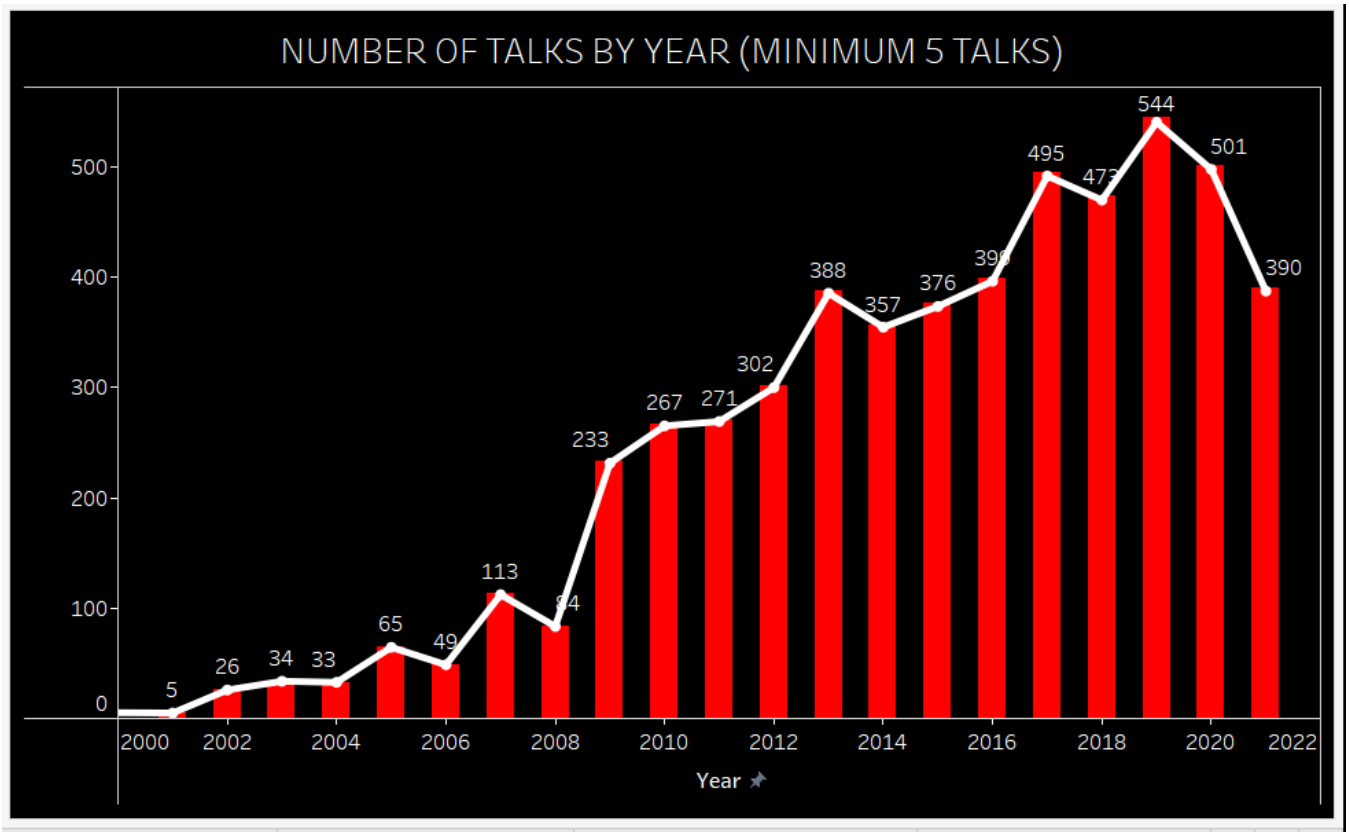
2 DATA VISUALIZATION

2.1 INTRODUCTION

Tableau is a powerful data visualization tool that enables us to create interactive and visually compelling visualizations. In this document, we will explore how to leverage Tableau to visualize the TED Talks dataset. Tableau offers a user-friendly interface and a wide range of visualization options, including bar charts, line charts, scatter plots, and more. By harnessing Tableau's capabilities, we can effectively analyze and present data in an engaging and informative manner. Through step-by-step instructions and examples, we will demonstrate how to create dynamic visualizations that uncover insights within the TED Talks dataset. By mastering data visualization with Tableau, we can effectively communicate findings, identify trends, and make data-driven decisions.

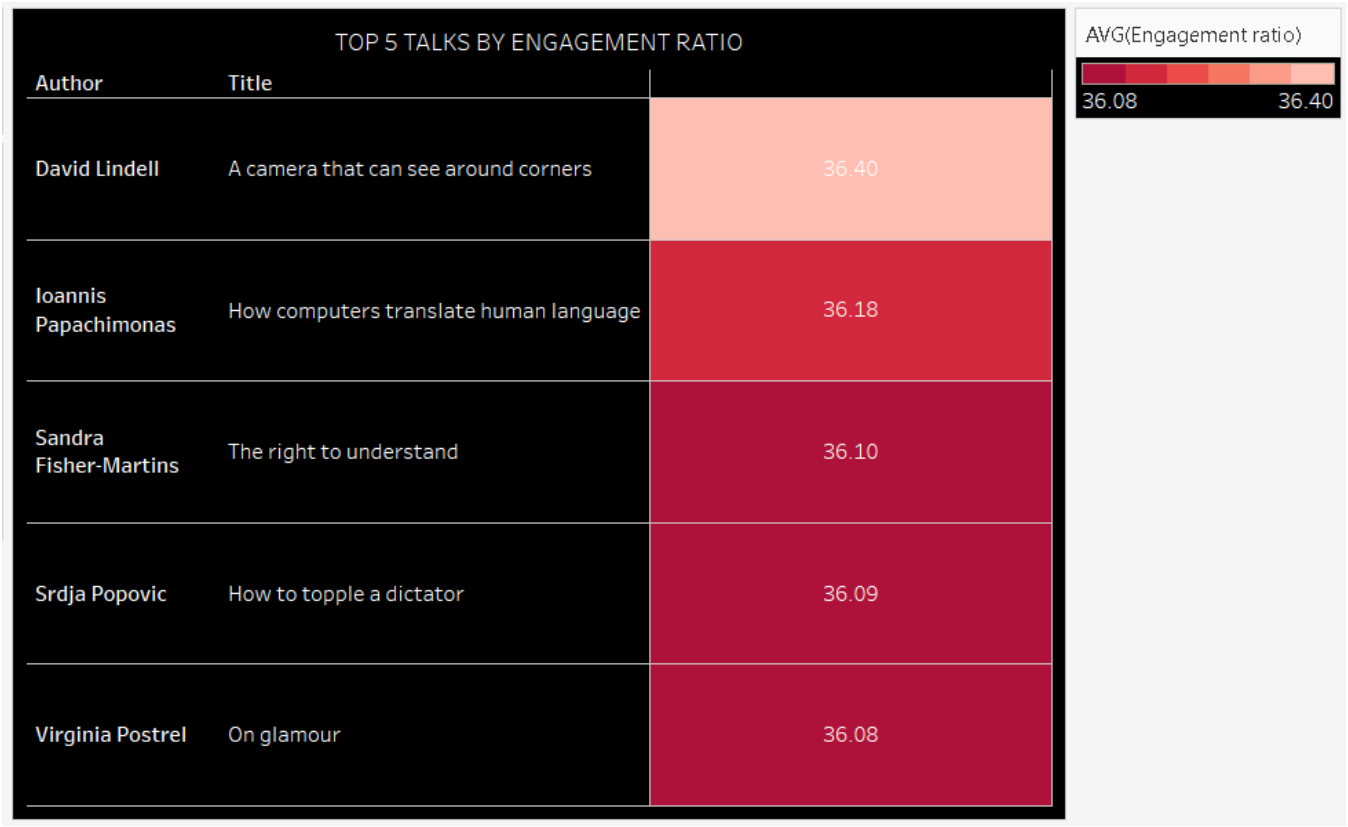
2.2 NUMBER OF TALKS BY YEAR

In this analysis, we explore the trend of the number of TED Talks delivered each year. Visualizations in the form of a line chart and a bar chart illustrate the increasing trend in the number of talks over time. By examining these charts, we can observe the growth and impact of TED Talks as a platform for sharing ideas and insights.



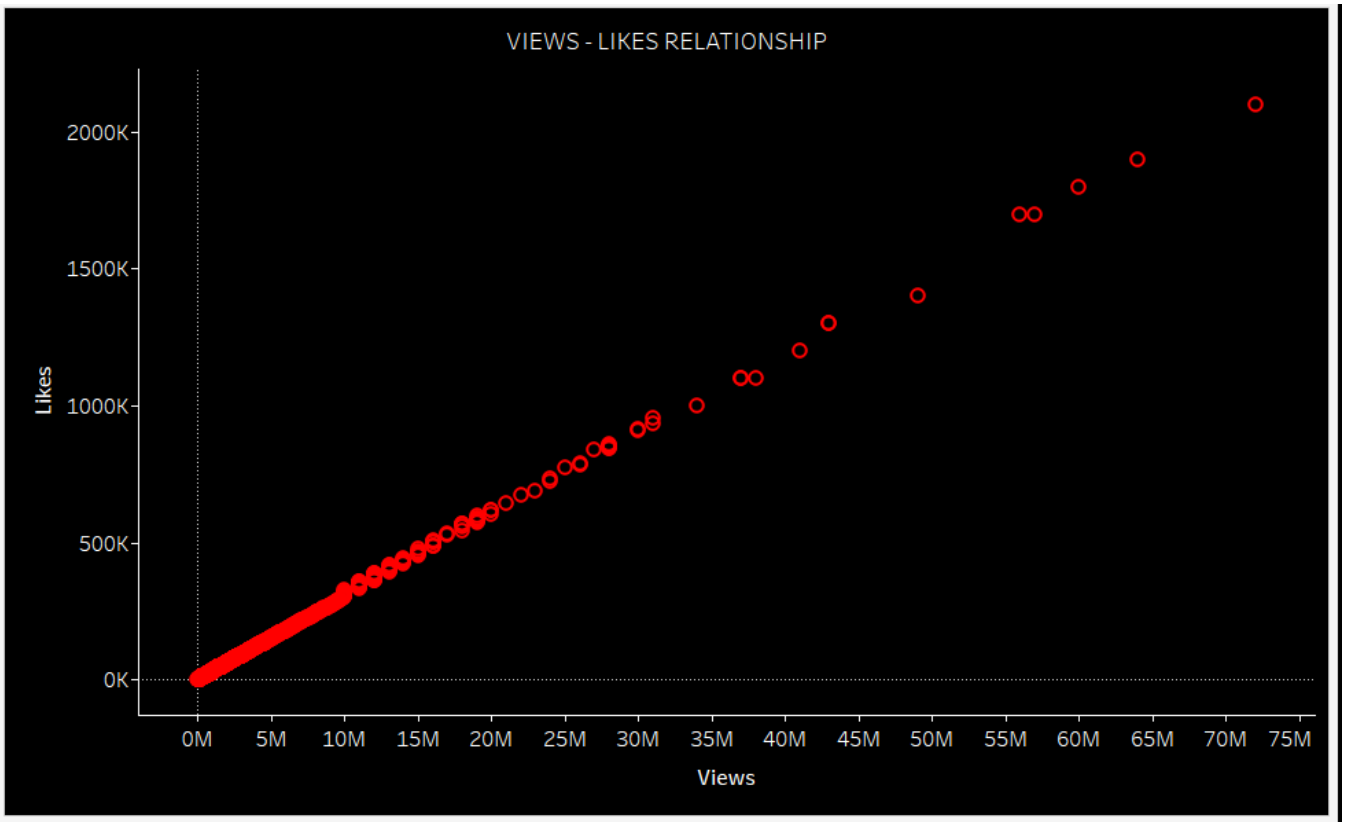
2.3 TOP 5 TED TALKS BY ENGAGEMENT RATIO

In this analysis, we identify the top 5 TED Talks with the highest engagement ratios. The engagement ratio is calculated by dividing the number of views by the number of likes for each talk. To present this information, we construct a table using highlight tables, highlighting the top-performing talks based on their engagement ratios. This analysis provides insights into the talks that generate the most audience interest and interaction within the TED Talks community.



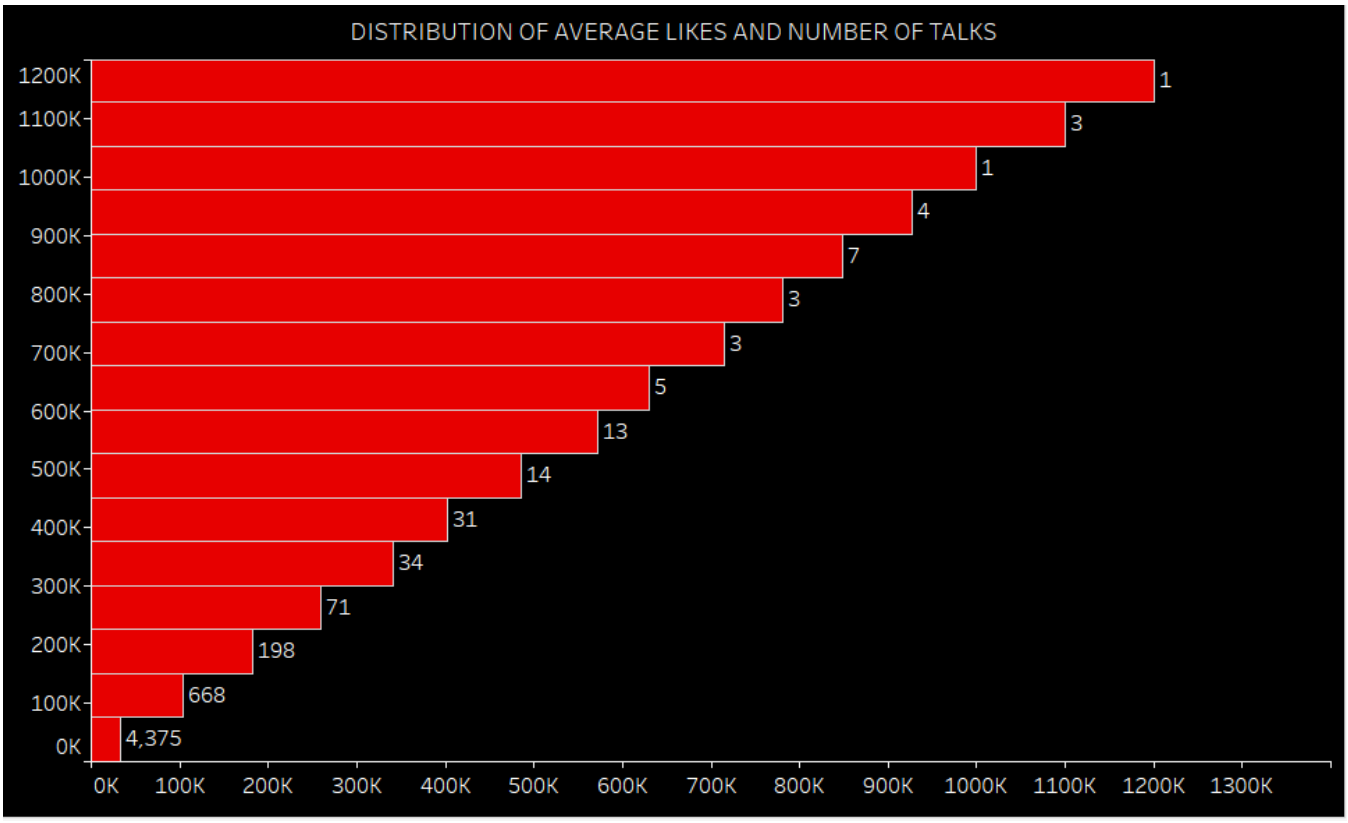
2.4 VIEWS - LIKES RELATIONSHIP

In this analysis, we explore the relationship between the number of views and likes for TED Talks. Using a scatter plot, we visualize how the number of likes corresponds to the number of views. This analysis helps us understand the level of audience engagement and whether an increase in views is associated with an increase in likes. By examining the scatter plot, we can gain insights into the popularity and impact of TED Talks within the community.



2.5 DISTRIBUTION OF AVERAGE LIKES AND NUMBER OF TALKS

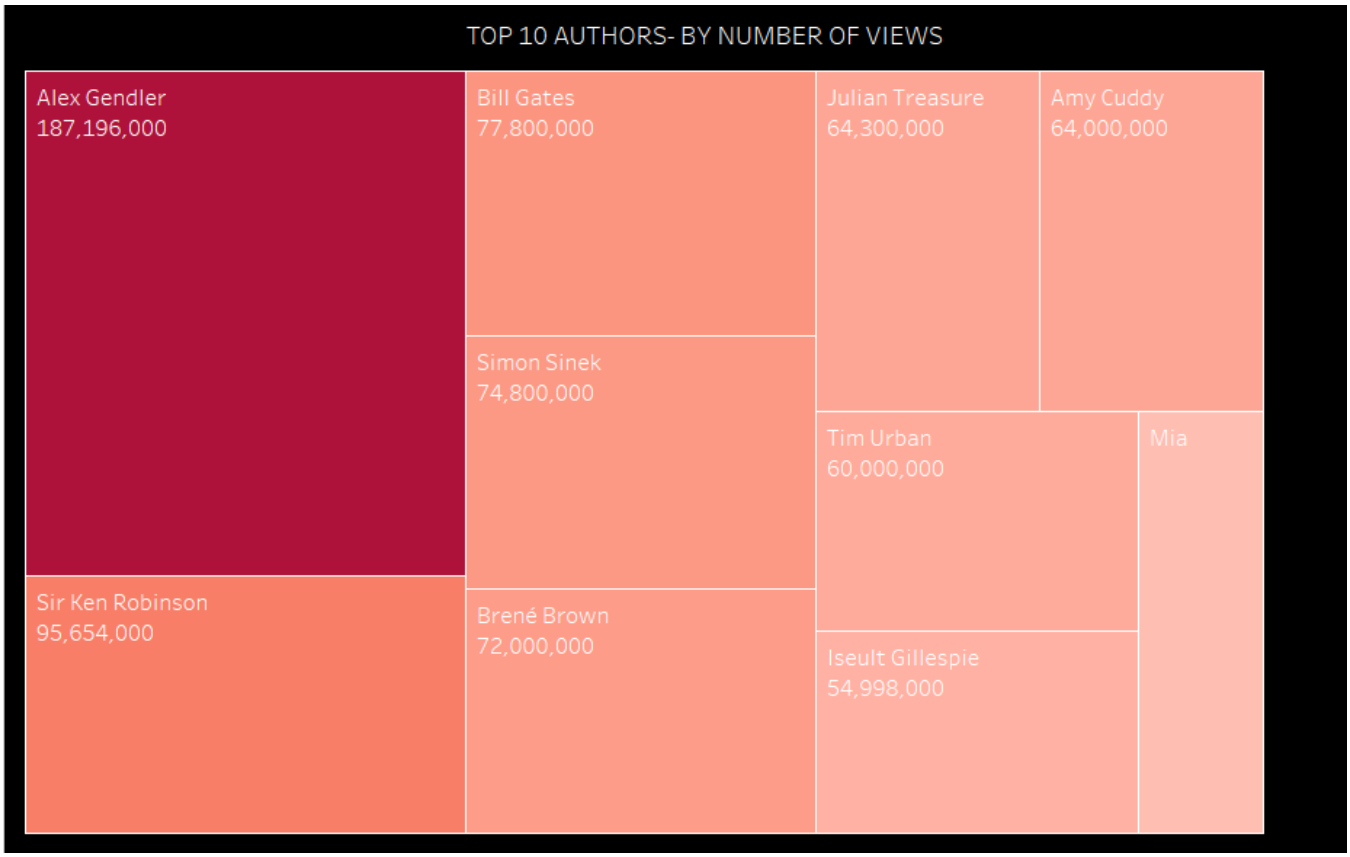
In this analysis, we visualize the distribution of average likes and the number of talks in the TED Talks dataset using a histogram. By examining the histogram, we can gain insights into the concentration and dispersion of data points, allowing us to understand the patterns and trends in the engagement levels and popularity of talks within the TED Talks community.



2.6 Top 10 Authors

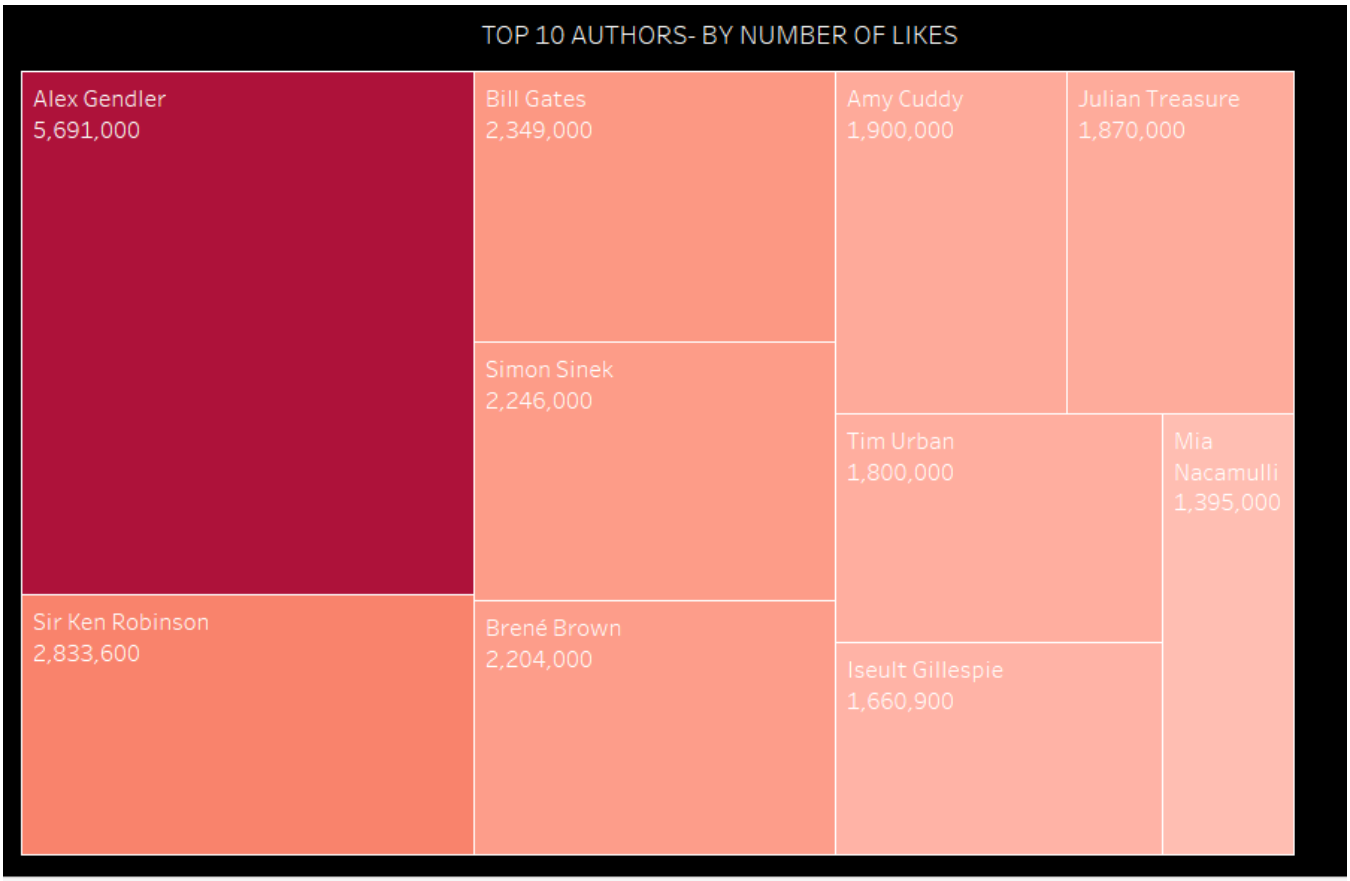
2.6.1 By Number of Views

In this analysis, we visualize the top 10 authors in the TED Talks dataset based on the number of views they have received. Using a treemap visualization, we present the authors as rectangles, with larger rectangles representing authors with higher view counts. This provides a clear visual representation of the most viewed authors and highlights their impact and popularity within the TED Talks community.



2.6.2 By Number of Likes

In this analysis, we visualize the top 10 authors in the TED Talks dataset based on the number of likes they have received. Using a treemap visualization, we represent the authors as rectangles, where the size of each rectangle corresponds to the number of likes received. This allows us to easily identify the most liked authors and emphasizes their influence and the appreciation for their talks within the TED Talks community.



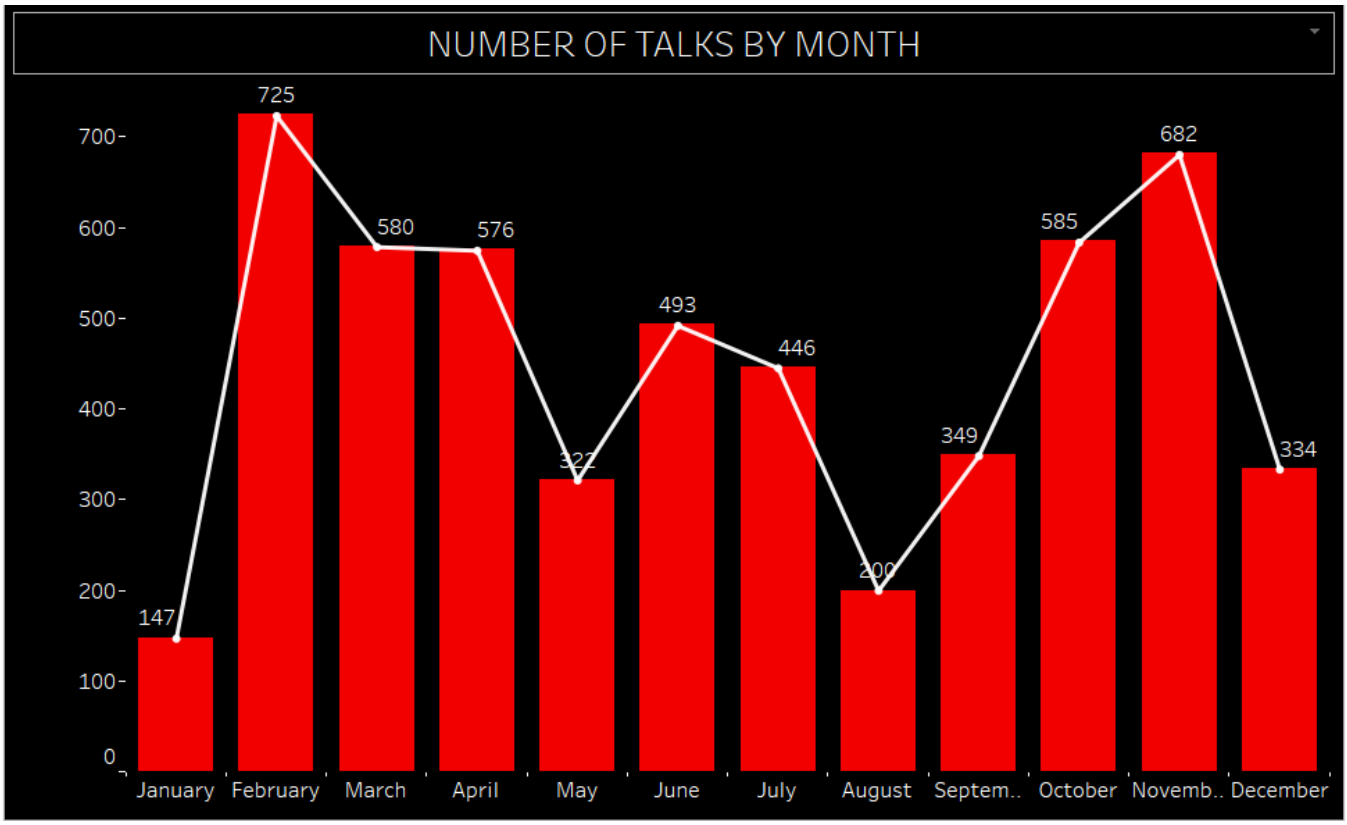
Upon comparing the top 10 authors by views and the top 10 authors by likes, we observe a significant overlap between the two sets. This indicates that the authors who receive a high number of views also tend to accumulate a considerable number of likes. This strong correlation suggests that popular talks with a large viewership are often well-received by the audience, leading to a higher number of likes.

The shared authors in both treemap visualizations highlight their exceptional influence and impact within the TED Talks community. Their talks not only attract a wide viewership but also resonate strongly with the audience, resulting in a significant number of likes.

Overall, this comparison underscores the close relationship between the number of views and likes, emphasizing the popularity and engagement of top authors within the TED Talks platform.

2.7 NUMBER OF TALKS BY MONTH

In this analysis, we visualize the distribution of the number of talks by month in the TED Talks dataset using both a bar chart and a line chart. These visualizations provide insights into the patterns and trends in talk counts over the months. The bar chart displays the number of talks for each month, allowing for easy comparison and identification of months with higher or lower activity. The line chart illustrates the progression of talk counts over time, enabling the observation of any trends or fluctuations in the distribution. Through these visualizations, we gain an understanding of the popularity and seasonality of talks within the TED Talks community. These charts help us identify any patterns related to the timing of TED Talk events and provide insights into the distribution of talks throughout the year.

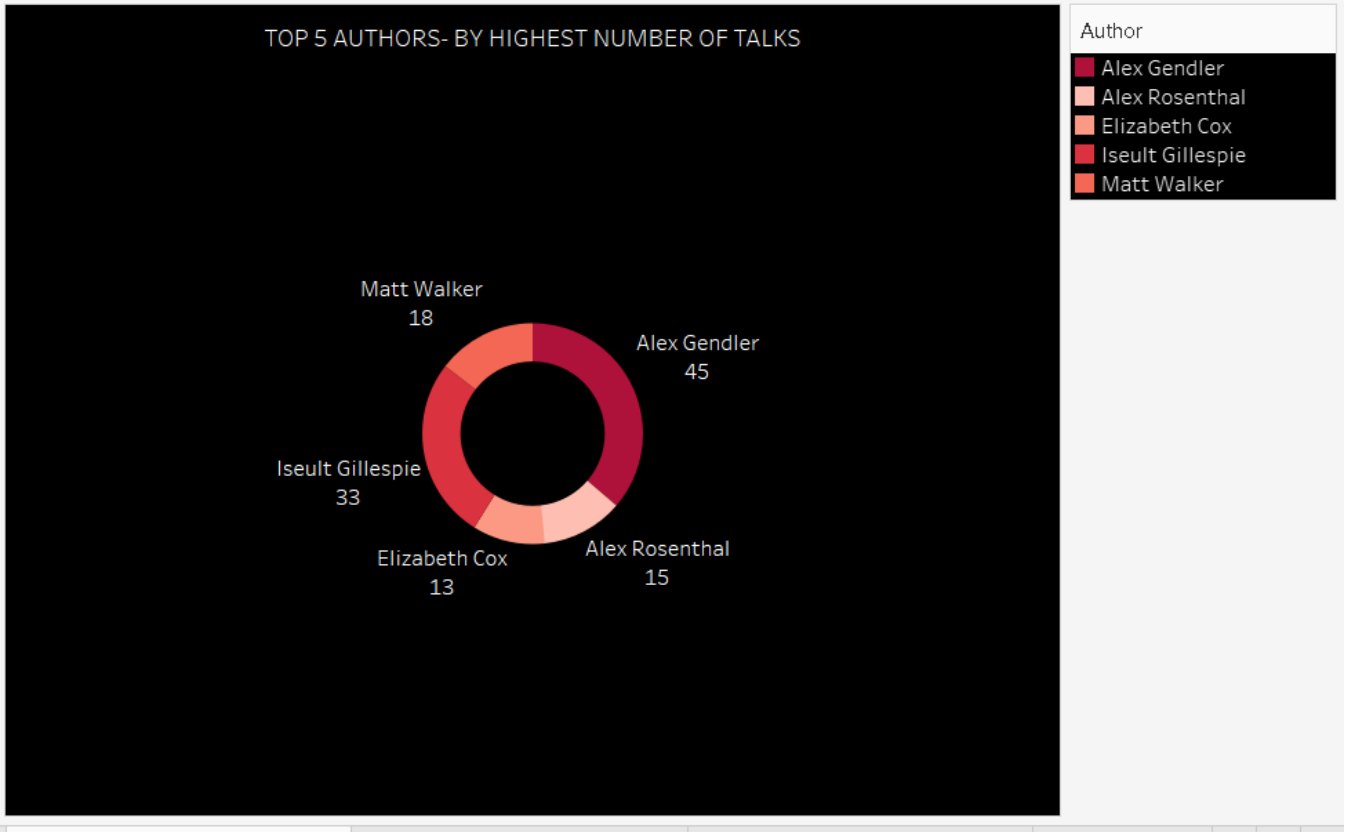


2.8 TOP 5 AUTHORS- BY NUMBER OF TALKS

In this analysis, we identify the top 5 authors with the highest number of talks in the TED Talks dataset. Using a donut chart visualization, we showcase the distribution of talks among these authors. The size of each segment in the donut chart corresponds to the number of talks delivered by each author.

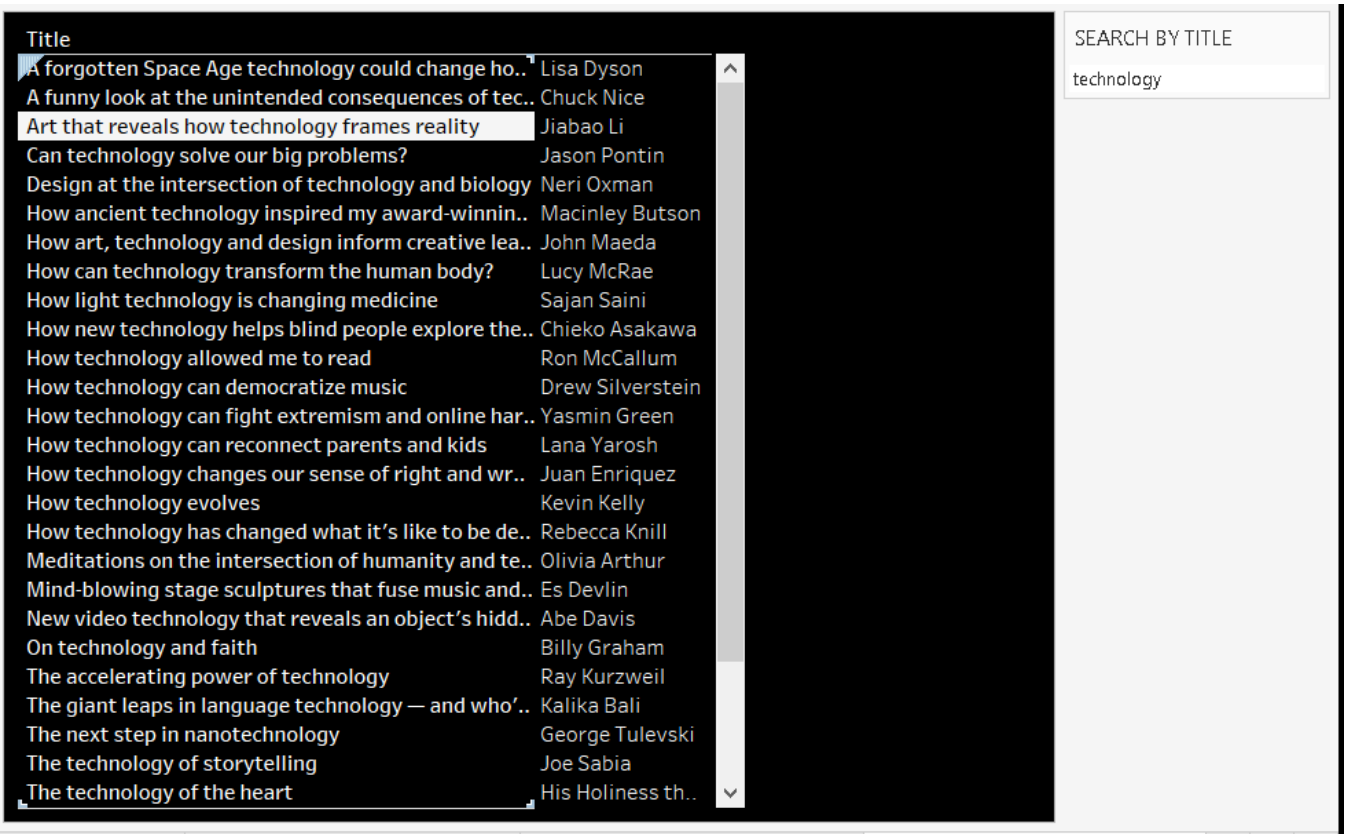
Alex Gendler emerges as the author with the highest number of talks, followed by Iseult Gillespie, Matt Walker, and others. This analysis highlights the significant contributions of these authors to the TED Talks community, as they have delivered multiple talks on various topics.

Through this analysis, we gain insights into the top authors by talk count, emphasizing their influence and impact within the TED Talks platform.

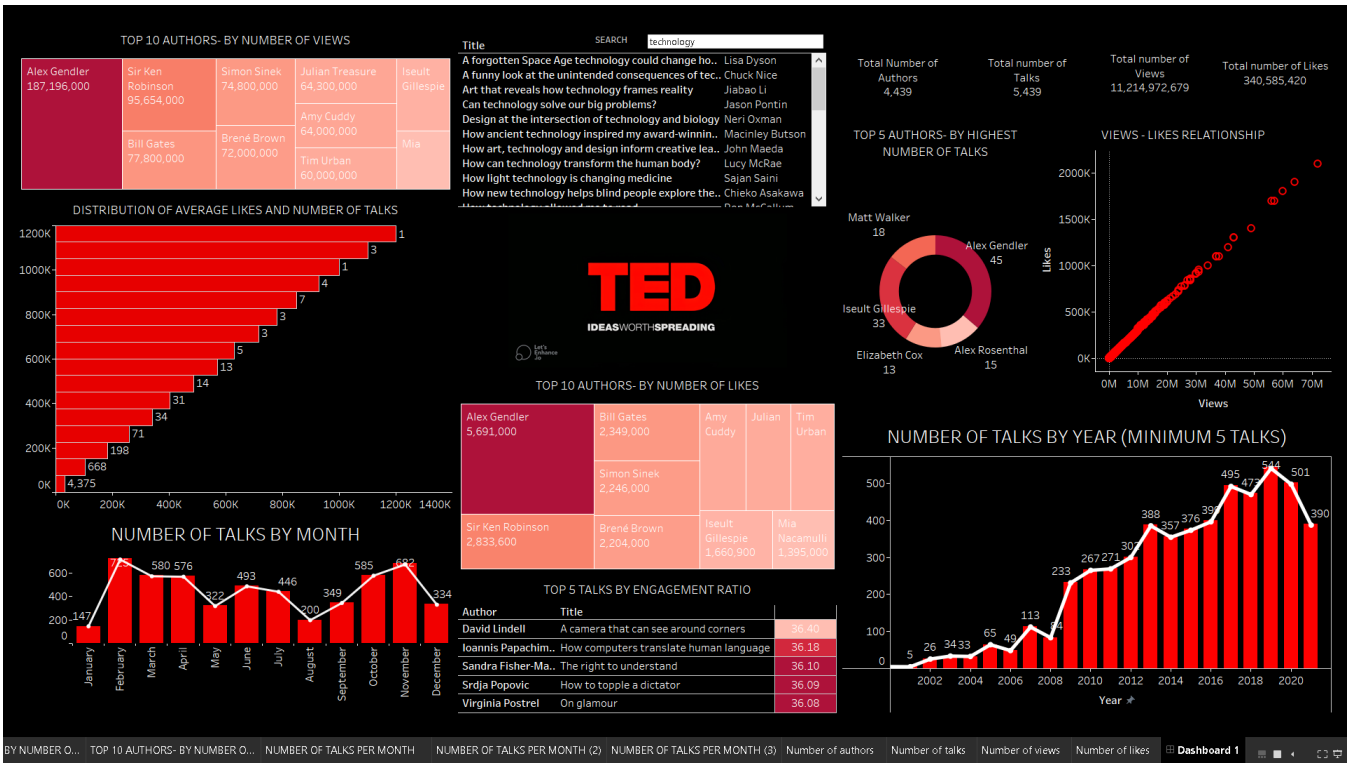


2.9 SEARCH BY KEYWORD

In this analysis, we provide a search functionality to help users find TED Talks by entering a keyword. The resulting chart lists all talks that contain the keyword in their title, allowing users to conveniently discover talks related to specific topics or interests. This search feature offers an efficient way to explore relevant talks within the TED Talks dataset and access valuable insights shared by speakers in the TED Talks community.



2.10 DASHBOARD - TED TALKS



2.11 INSIGHTS

1. Talks have been increasing over time, indicating the growing popularity of TED Talks.
2. There is a positive correlation between views and likes, suggesting higher engagement for popular talks.
3. Top authors by views and likes largely overlap, indicating consistent popularity and appreciation.
4. Talk counts vary across months, revealing seasonal patterns and trends.
5. Alex Gendler has the highest number of talks, followed by Iseult Gillespie and Matt Walker.
6. Keyword search enables users to find talks based on specific interests or topics.

Overall, the analysis reveals the growth and impact of TED Talks, the relationship between views and likes, the influence of top authors, seasonal trends, and the ability to search for talks based on keywords.

CONCLUSION

The combined analysis using pandas and visualization with Tableau provided valuable insights into the TED Talks dataset. We uncovered trends such as the increasing number of talks over time, the correlation between views and likes, and the distribution of talks by month. The visualizations effectively communicated these insights, highlighting the impact of top authors, seasonal patterns, and the ability to search for talks based on keywords. Overall, this analysis enhanced our understanding of the TED Talks community, showcasing its growth, engagement, and the diverse range of topics covered.