project overview

This dataset is created for beginner students of data analysis who can

explore the field with real-life data. using TED talk data will help

them to analyse the talks and they can also watch the talks of their

favourite author with the help of the dataset as well

- **what to do in this project**
  - Finding the most popular TED talks
  - finding the most popular ted talks speakers (in terms of number of talks)
  - month wise analysis of ted talk frequency
  - year wise analysis of ted talk frequency
  - finding ted talks of your favourite author
  - finding ted talks with the best view to like ratio
  - finding ted talks based on tags(like climate)
  - finding the most popular ted talks speaker (in terms of number of views)

In [1]:
```
# Importing and reading the file
```

In [2]:
```
import pandas as pd
import plotly.express as px
import numpy as np

# reading
df = pd.read_csv('ted_data.csv')
df.head()
```

Out[2]:

	title	author	date	views	likes	link
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	December 2021	404000	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	February 2022	214000	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	How play can spark new ideas for your business	Martin Reeves	September 2021	412000	12000	https://ted.com/talks/martin_reeves_how_play_c...
3	Why is China appointing judges to combat clima...	James K. Thornton	October 2021	427000	12000	https://ted.com/talks/james_k_thornton_why_is_...
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	October 2021	2400	72	https://ted.com/talks/mahendra_singhi_cement_s...

In [3]:
```
# describing the dataframe
df.describe()
```

Out[3]:

	views	likes
count	5.440000e+03	5.440000e+03
mean	2.061576e+06	6.260762e+04
std	3.567098e+06	1.076468e+05
min	5.320000e+02	1.500000e+01
25%	6.707500e+05	2.000000e+04
50%	1.300000e+06	4.050000e+04
75%	2.100000e+06	6.500000e+04
max	7.200000e+07	2.100000e+06

In [4]:
```
# getting the information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5440 entries, 0 to 5439
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   title   5440 non-null   object
 1   author  5439 non-null   object
 2   date    5440 non-null   object
 3   views   5440 non-null   int64 
 4   likes   5440 non-null   int64 
 5   link    5440 non-null   object
dtypes: int64(2), object(4)
memory usage: 255.1+ KB
```

In [5]:
```
# checking the null values
df.isnull().sum()
```

Out[5]:
```
title     0
author    1
date      0
views     0
likes     0
link      0
dtype: int64
```

In [6]:
```
# filling the null values using the backward fill
df = df.fillna(method='bfill')
```

In [7]:
```
# pre -processing the date column
```

In [8]:
```
year = []
month = []
for i in df['date']:
    month.append(i.split(' ')[0])
    year.append(i.split(' ')[1])
```

In [9]:
```
df['month'] = month
df['year'] = year
del df['date']
df.head()
```

Out[9]:

	title	author	views	likes	link	month	year
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	404000	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...	December	2021
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	214000	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...	February	2022
2	How play can spark new ideas for your business	Martin Reeves	412000	12000	https://ted.com/talks/martin_reeves_how_play_c...	September	2021
3	Why is China appointing judges to combat clima...	James K. Thornton	427000	12000	https://ted.com/talks/james_k_thornton_why_is_...	October	2021
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2400	72	https://ted.com/talks/mahendra_singhi_cement_s...	October	2021

In [10]:
```
# adding the views to like ratio column
v_l = []
for i in df.values:
    v_l.append(round(i[2]/i[3],2))

df['views_to_likes'] = v_l
```

In [11]:
```
# finding the most popular ted talks
speaker = {}
for i in df.values:
    if i[1] not in speaker:
        speaker[i[1]] = 1
    else:
        speaker[i[1]] += 1


most_popular_speaker = pd.DataFrame()
most_popular_speaker['Speaker'] = speaker.keys()
most_popular_speaker['total_view'] = speaker.values()
most_popular_speaker.sort_values(by='total_view',ascending = False).h
```

Out[11]:

	Speaker	total_view
63	Alex Gendler	45

In [12]:
```
# finding the most popular ted talks speakers (in terms of number of
speaker = {}
for i in df.values:
    if i[1] not in speaker:
        speaker[i[1]] = 1
    else:
        speaker[i[1]] += 1

most_popular                    = pd.DataFrame()    # creating the dat
most_popular['speaker_name']    = speaker.keys()    # adding column
most_popular['times_appeared'] = speaker.values()  # adding another c
most_popular.sort_values(by='times_appeared',ascending = False).head(
```

Out[12]:

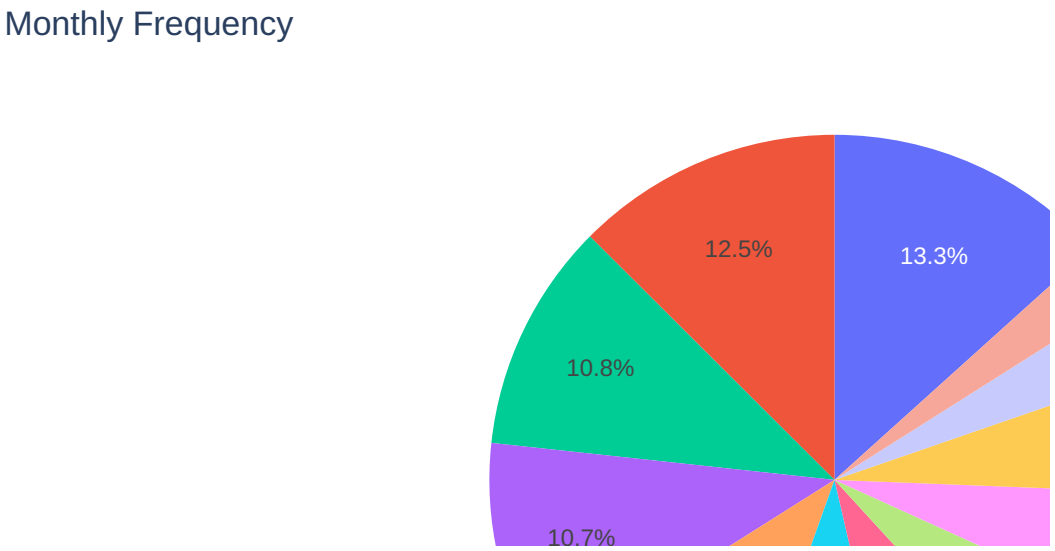
	speaker_name	times_appeared
63	Alex Gendler	45

In [13]:
```
# month wise analysis of ted talk frequency
freq_m = {'January':0,
          'February':0,
          'March':0,
          'April':0,
          'May':0,
          'June':0,
          'July':0,
          'August':0,
          'September':0,
          'October':0,
          'November':0,
          'December':0,
         }

for i in df.values:
    if i[5]  not in freq_m:
        freq_m[i[5]] = 0
    else:
        freq_m[i[5]] += 1

month = pd.DataFrame()
month['months'] = freq_m.keys()
month['Frequency'] = freq_m.values()
month.head()
```

Out[13]:

	months	Frequency
0	January	147
1	February	725
2	March	580
3	April	576
4	May	322

In [14]:
```
fig = px.pie(month,values ='Frequency',names = 'months',title = 'Mont
fig.show()
```

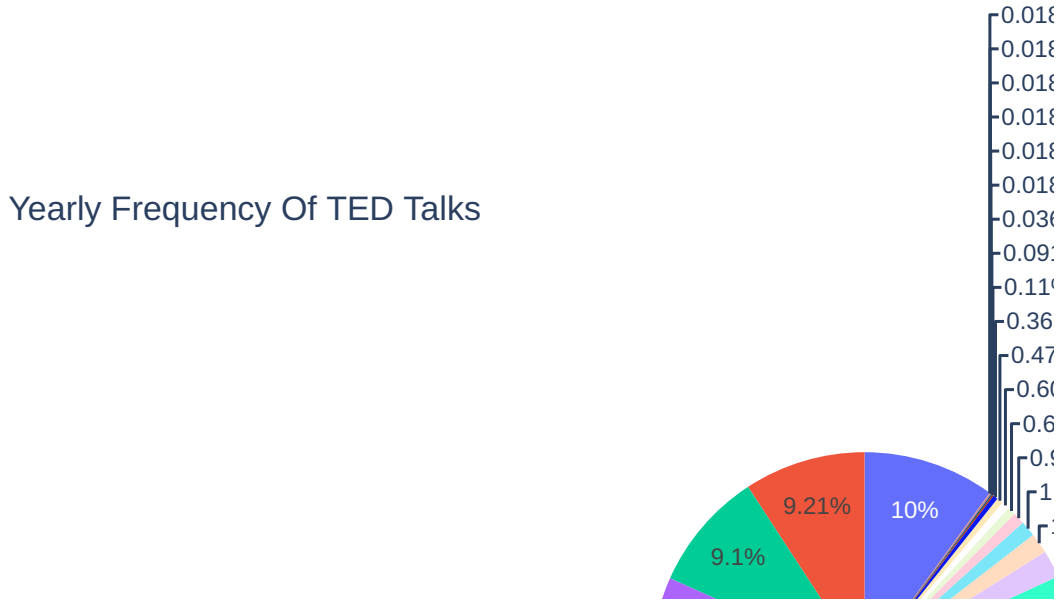


In [15]:
```
# year wise analysis of ted talk frequency

freq = {}
for i in df.values:
    if i[6]  not in freq:
        freq[i[6]] = 1
    else:
        freq[i[6]] += 1

year = pd.DataFrame()
year['Year'] = freq.keys()
year['Frequency'] = freq.values()
year = year.sort_values(by = 'Year', ascending = True)
year.head()
```

Out[15]:

	Year	Frequency
9	1970	2
19	1972	1
22	1983	1
29	1984	1
28	1990	1

In [16]:
```
fig = px.pie(year,values ='Frequency',names = 'Year',title = 'Yearly
fig.show()
```



In [17]:
```
# finding ted talks of your favourite author
author =  input('Enter the name of the author')
for i in df.values:
    if author.lower() == i[1].lower():
        print(i[0])
```

```
Enter the name of the authorMing Luke
What's the point(e) of ballet?
What's a squillo, and why do opera singers need it?
```

In [18]:
```
# finding ted talks with the best view to like ratio
df[df['views_to_likes'] == max(df['views_to_likes'])]
```

Out[18]:

	title	author	views	likes	link	month	year	views_to_
837	How to see more and care less: The art of Geor...	Iseult Gillespie	364000	10000	https://ted.com/talks/iseult_gillespie_how_to_...	June	2020	
905	What's the point(e) of ballet?	Ming Luke	364000	10000	https://ted.com/talks/ming_luke_what_s_the_poi...	April	2020	
955	A camera that can see around corners	David Lindell	364000	10000	https://ted.com/talks/david_lindell_a_camera_t...	November	2019	

In [45]:
```
# finding ted talks based on tags(like climate)

# recommending based on tags
tags = input('Enter the keyword : ')
c = 1
for i in df.values:
    if tags.capitalize() in i[0].split(' ') or tags.lower() in i[0].s
        print(str(c) + ') ' + i[0])
        c += 1
```

```
Enter the keyword : India
1) Why was India split into two countries?
2) How India could pull off the world's most ambitious energy trans
ition
3) How the coronavirus is impacting India — and what needs to happe
n next
4) 5 steps for clean air in India
5) A bold plan to empower 1.6 million out-of-school girls in India
6) How women in rural India turned courage into capital
7) The rise of cricket, the rise of India
```

In [20]:
```
# finding the most popular ted talks speaker (in terms of number of v
speak = {}
for i in df.values:
    if i[1] not in speak:
        speak[i[1]] = i[2]
    else:
        speak[i[1]] += i[2]


most_viewed_speaker = pd.DataFrame()
most_viewed_speaker['Speaker'] = speak.keys()
most_viewed_speaker['total_view'] = speak.values()
most_viewed_speaker.sort_values(by='total_view',ascending = False).he
```

Out[20]:

	Speaker	total_view
63	Alex Gendler	187196000

In [21]:
```
# another way to approach the above problem is
new = df.groupby(by = 'author').sum()
new[new['views'] == max(new['views'])]
```

```
/tmp/ipykernel_4635/1133485735.py:2: FutureWarning:

The default value of numeric_only in DataFrameGroupBy.sum is deprec
ated. In a future version, numeric_only will default to False. Eith
er specify numeric_only or select only columns which should be vali
d for the function.
```

Out[21]:

	views	likes	views_to_likes
author			
Alex Gendler	187196000	5691000	1494.1