```
#Data analysis on K.J YouTuber
Dan Hua Li
12/7/2022

This project attempt to help better understand the growth of YouTube
community and find the best ways that as a data scientist how to improve the benefit
for their Youtuber in data science area. Definitely, the best ways to be a successful
YouTuber is to tell a wonderful story and work hard on the content. However, here I
just ignore the content of Ken Jee's YouTube but focus on the title, video's length and I
assume those are the important factors that impact the profit. Hopefully, through this
project may help making profit on our channel at the beginning.
This project objective will try to answer the follow questions:
• what is the outlook of data science, is it popular?
• What types of video titles drive the most traffic?
• Does it exist a appropriate length of video that could help maximizing the
profit?
This project will combine the tools of Excel, Python and Weka, to envision, execute,
and summarize above issues based on a data-science-oriented study. Processing data
includes:
▪ Data description
- Data Background information
- Data Dictionary
- Missing values
▪ Data mining process
- data cleansing,
- attribute selection
- transformation,
- training and testing process (10/5-fold cross-validation).
- Linear Regression model
▪ Final Results and Recommendation
Data Description
 Data Background information: The data for this project is loading from Ken's
Kaggle, a famous YouTuber in data science, who provided his personal YouTube data
Dan Hua Li
for analysis. Notice: the study is only based on the data science YouTuber and the
limitation of Ken Jee's Private YouTube Data source. The data set I selected includes
two parts:
1) Aggregated Metrics By Video with Country and Subscriber Status
- data includes dimensions for which country people are viewing from and if
the viewers are subscribed to the channel or not. - Attributes:15 instances:
55292
2) Aggregated Metrics By Video
- includes all the topline metrics from the channel from its start (around 2015 to
Jan 22 2022). There are 111857 original records and group it into 224 records.
```

In [2]:

```python
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

In [3]:

```python
# Using 1st data to analyze videos in geographical way
Country_df = pd.read_csv('Data_Aggregated_Metrics_By_Country_And_Subscriber_Status.csv')
Country_df
```

```
             Explained #...

         ➶ Hot Topics in
      2       Tech: Data  OtqQYqRNDGI   59  https://i.ytimg.com/vi/OtqQYqRNDGI/hqdefault.jpg   RW    True     2    0    0    0
                 Science
             Explained #...

         ➶ Hot Topics in
      3       Tech: Data  OtqQYqRNDGI   59  https://i.ytimg.com/vi/OtqQYqRNDGI/hqdefault.jpg   US    True   979   81    6    8
                 Science
             Explained #...

         ➶ Hot Topics in
      4       Tech: Data  OtqQYqRNDGI   59  https://i.ytimg.com/vi/OtqQYqRNDGI/hqdefault.jpg   DE   False    75    3    0    1
                 Science
             Explained #...

    ...              ...          ...  ...                                               ...  ...     ...   ...  ...  ...  ...

           #66DaysOfData
  55287   - 3 Reasons to  sICJ6a2wX5g   53  https://i.ytimg.com/vi/sICJ6a2wX5g/hqdefault.jpg   MM   False     1    0    0    0
                  Start!

           #66DaysOfData
```

In [4]:

```
pip install pycountry
```

```
Requirement already satisfied: pycountry in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (22.3.5)Note:
you may need to restart the kernel to use updated packages.

Requirement already satisfied: setuptools in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from pycoun
try) (63.2.0)
```

In [5]:

```
import pycountry
def do_fuzzy_search(country):
    try:
        result = pycountry.countries.search_fuzzy(country)
    except Exception:
        return np.nan
    else:
        return result[0].alpha_3

iso_map = {country: do_fuzzy_search(country) for country in Country_df["Country Code"].unique()}

Country_df["Country_Code"] = Country_df["Country Code"].map(iso_map)

Country_df = Country_df.loc[~(Country_df['Country_Code'].isna()),]

GIS_plot_df = Country_df.groupby(by=['Country_Code', 'Country Code'], as_index=False, dropna=True).mean()
GIS_plot_df.head()
```

Out[5]:

	Country_Code	Country Code	Video Length	Is Subscribed	Views	Video Likes Added	Video Dislikes Added	Video Likes Removed	User Subscriptions Added	User Subscriptions Removed	Average View Percentage	Average Watch Time	User Comments Added
0	ABW	AW	700.216867	0.361446	3.036145	0.048193	0.000000	0.000000	0.132530	0.000000	0.394759	200.441195	0.0
1	AFG	AF	746.267857	0.422619	5.160714	0.184524	0.011905	0.011905	0.202381	0.017857	0.268558	137.758473	0.0
2	AGO	AO	912.624204	0.401274	5.210191	0.184713	0.012739	0.000000	0.171975	0.006369	0.333852	193.719285	0.0
3	ALA	AX	490.285714	0.000000	1.857143	0.000000	0.000000	0.000000	0.000000	0.000000	0.591437	333.560914	0.0
4	ALB	AL	906.320423	0.461268	9.728873	0.274648	0.035211	0.014085	0.123239	0.010563	0.240977	150.988097	0.0

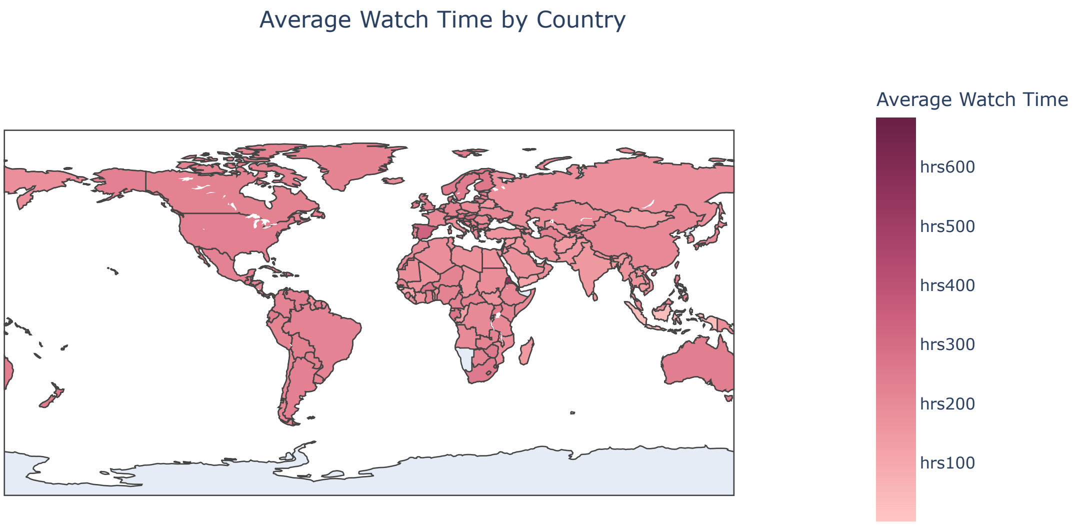
In [6]:

```
pip install plotly
```

```
Requirement already satisfied: plotly in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (5.11.0)Note: yo
u may need to restart the kernel to use updated packages.

Requirement already satisfied: tenacity>=6.2.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from p
lotly) (8.1.0)
```

In [7]:

```
import plotly.graph_objects as go
fig = go.Figure(data=go.Choropleth(locations = GIS_plot_df['Country_Code'],
                                   z = GIS_plot_df['Average Watch Time'],
                                   text = GIS_plot_df['Country Code'],
                                   colorscale = 'burg',
                                   colorbar_tickprefix = 'hrs',
                                   colorbar_title = 'Average Watch Time')
)
fig.update_layout(
    title={'text': 'Average Watch Time by Country',
           'y':0.9,
           'x':0.5,
           'xanchor': 'center',
           'yanchor': 'top'})

fig.show()
```



In [27]:

```
import seaborn as sns
import os
```

In [28]:

```
# check the sum of null values
Country_df.isnull().sum()
```

Out[28]:

```
Video Title                      0
External Video ID                0
Video Length                     0
Thumbnail link                   0
Country Code                   386
Is Subscribed                    0
Views                            0
Video Likes Added                0
Video Dislikes Added             0
Video Likes Removed              0
User Subscriptions Added         0
User Subscriptions Removed       0
Average View Percentage       1438
Average Watch Time            1438
User Comments Added              0
dtype: int64
```

In [31]:

```
import numpy as np # linear algebra
import pandas as pd
import seaborn as sns
import os
```

In [34]:
```
#Aggregated_Metrics_By_Video2
video_df=pd.read_csv('Aggregated_Metrics_By_Video.csv')
video_df
```

Out[34]:

	Video	Video title	Video publish time	Comments added	Shares	Dislikes	Likes	Subscribers lost	Subscribers gained	RPM (USD)	CPM (USD)	Average percentage viewed (%)	Average view duration	Views	W
0	Total	NaN	NaN	14197	39640	3902	225021	45790	229241	5.276	11.990	26.61	0:03:25	5568487	31
1	4OZip0cgOho	How I Would Learn Data Science (If I Had to St...	8-May-20	907	9583	942	46903	451	46904	6.353	12.835	36.65	0:03:09	1253559	65
2	78LjdAAw0wA	100K Channel Update + AMA Stream!	12-Nov-20	412	4	4	130	15	12	2.668	6.259	6.26	0:05:14	2291	
3	hO_YKK_0Qck	Uber Driver to Machine Learning Engineer in 9 ...	16-Jul-20	402	152	15	881	9	198	9.516	11.695	15.12	0:10:21	21350	3
4	uXLnbdHMf8w	Why I'm Starting Data Science Over Again.	29-Aug-20	375	367	22	2622	40	1957	3.143	7.943	33.41	0:02:36	49564	2

Missing Values --> Data Cleansing

In [35]:
```
video_df=pd.read_csv('Aggregated_Metrics_By_Video.csv')
video_df.isnull().sum()
```

Out[35]:
```
Video                                 0
Video title                           1
Video publish time                    1
Comments added                        0
Shares                                0
Dislikes                              0
Likes                                 0
Subscribers lost                      0
Subscribers gained                    0
RPM (USD)                             0
CPM (USD)                             2
Average percentage viewed (%)         0
Average view duration                 0
Views                                 0
Watch time (hours)                    0
Subscribers                           0
Your estimated revenue (USD)          0
Impressions                           0
Impressions click-through rate (%)    0
dtype: int64
```

In [40]:
```
video2_df= video_df.dropna()
video2_df.isnull().sum()
```

Out[40]:
```
Video                                 0
Video title                           0
Video publish time                    0
Comments added                        0
Shares                                0
Dislikes                              0
Likes                                 0
Subscribers lost                      0
Subscribers gained                    0
RPM (USD)                             0
CPM (USD)                             0
Average percentage viewed (%)         0
Average view duration                 0
Views                                 0
Watch time (hours)                    0
Subscribers                           0
Your estimated revenue (USD)          0
Impressions                           0
Impressions click-through rate (%)    0
dtype: int64
```

In [3]:

```
pip install statsmodels
```

```
Collecting statsmodelsNote: you may need to restart the kernel to use updated packages.

  Downloading statsmodels-0.13.5-cp310-cp310-win_amd64.whl (9.1 MB)
     ---------------------------------------- 9.1/9.1 MB 4.1 MB/s eta 0:00:00
Collecting patsy>=0.5.2
  Downloading patsy-0.5.3-py2.py3-none-any.whl (233 kB)
     -------------------------------------- 233.8/233.8 kB 1.6 MB/s eta 0:00:00
Requirement already satisfied: packaging>=21.3 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from s
tatsmodels) (21.3)
Requirement already satisfied: pandas>=0.25 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from stat
smodels) (1.4.4)
Requirement already satisfied: numpy>=1.22.3 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from sta
tsmodels) (1.23.2)
Requirement already satisfied: scipy>=1.3 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from statsm
odels) (1.9.1)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\yy\appdata\local\programs\python\python310\lib\site-package
s (from packaging>=21.3->statsmodels) (3.0.9)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages
(from pandas>=0.25->statsmodels) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from pand
as>=0.25->statsmodels) (2022.2.1)
Requirement already satisfied: six in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from patsy>=0.5.2-
>statsmodels) (1.16.0)
Installing collected packages: patsy, statsmodels
Successfully installed patsy-0.5.3 statsmodels-0.13.5
```

In [6]:

```
import pandas as pd
import statsmodels.api as sm
dataset = pd.read_csv('video view_like.csv')
dataset.head()
```

Out[6]:

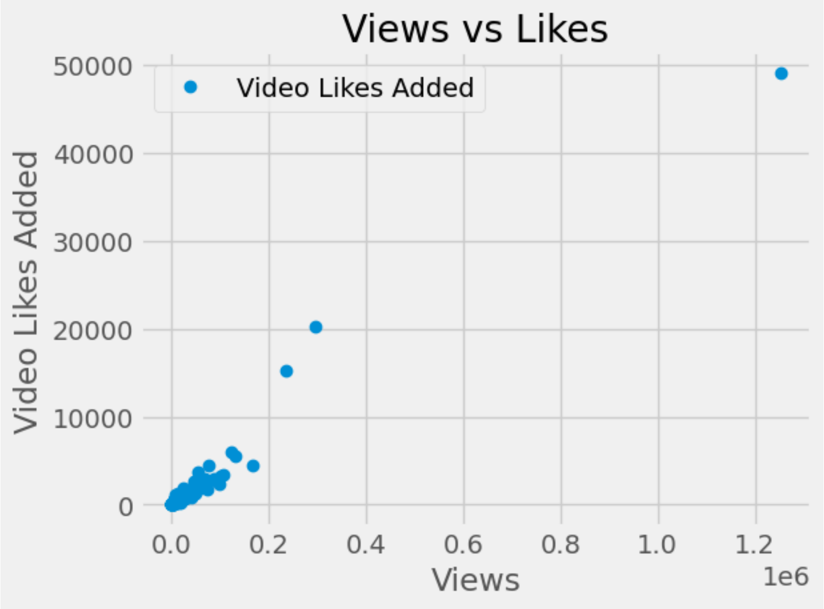
	Video Title	Video Length	Views	Video Likes Added
0	Hot Topics in Tech: Data Science Explained #SH...	59	8003	409
1	git for Data Science Made Simple... (Hopefully)	392	12629	667
2	Work From Home Data Scientist: Day in the Life	331	26582	754
3	Why is Balance Important in Data Science?	238	612	33
4	Why are APIs Important for Data Science?	322	6537	363

In [7]:

```
import matplotlib.pyplot as plt
%matplotlib inline
```

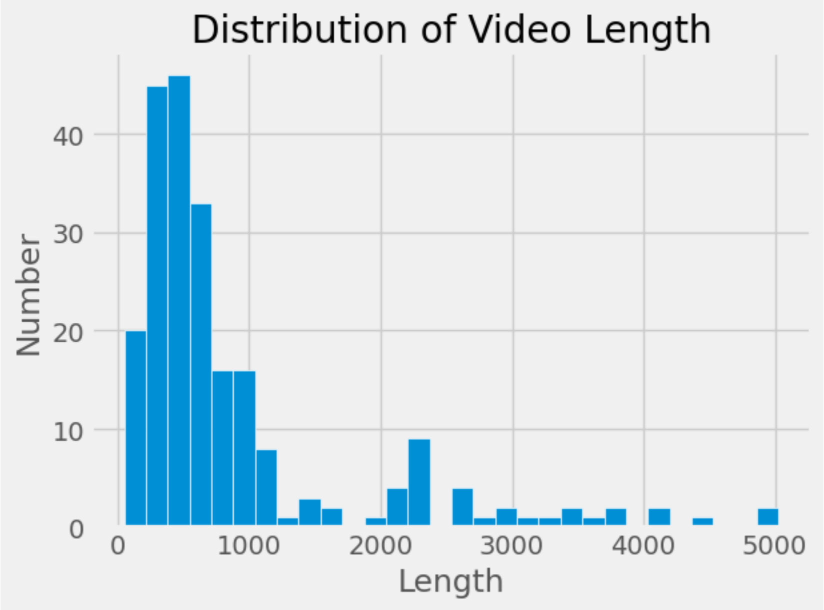
In [35]:

```
plt.style.use('fivethirtyeight')
dataset.plot(x='Views', y='Video Likes Added', style='o')
plt.title('Views vs Likes')
plt.xlabel('Views')
plt.ylabel('Video Likes Added')
plt.tight_layout()
plt.show()
```



In [ ]:

In [38]:
```
plt.hist(dataset["Video Length"],bins=30,edgecolor='white')
plt.title('Distribution of Video Length')
plt.xlabel('Length')
plt.ylabel('Number')
plt.tight_layout()
plt.show()
```



In [42]:
```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
%matplotlib inline
title = pd.read_csv('data top title of views.csv')
title.head()
```

Out[42]:

	Video Title	Video Length	Views	Video Likes Added	Unnamed: 4
0	How I Would Learn Data Science (If I Had to St...	516	1252970	49000	How I Would Learn Data Science (If I Had to St...
1	The Best Free Data Science Courses Nobody is T...	375	297050	20293	The Best Free Data Science Courses Nobody is T...
2	3 Proven Data Science Projects for Beginners (...	454	237192	15281	3 Proven Data Science Projects for Beginners (...
3	Beginner Kaggle Data Science Project Walk-Thro...	2296	167881	4523	Beginner Kaggle Data Science Project Walk-Thro...
4	The Projects You Should Do To Get A Data Scien...	770	131573	5458	The Projects You Should Do To Get A Data Scien...

In [47]:
```
#plt.bar(title['Video Title'], title['Views'])
#plt.title('most hot titles for view')
#plt.xlable('hot titles')
#plt.ylable('Number of views')
#plt.tight_layout()
#plt.show.head(5)
```

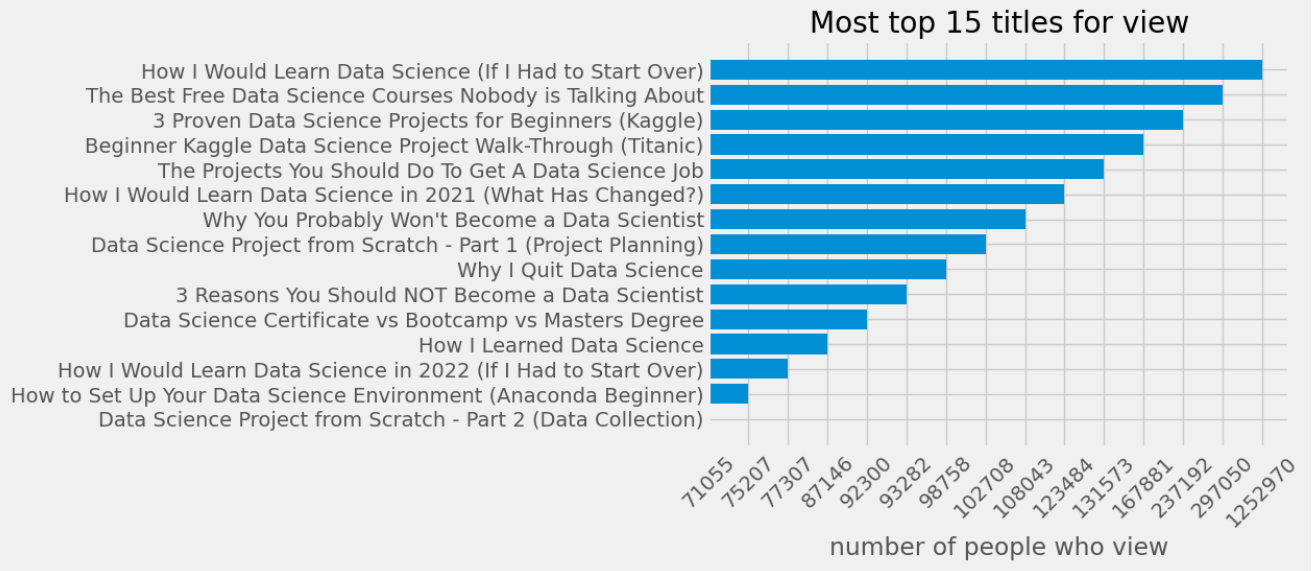
In [12]:
```
import csv
csvfile=open('data_title_views.csv','r')
data_title_views={}
for row in csv.DictReader(csvfile):
    #print(row)
    data_title_views[row['Title']]=row['Views']
    #print(data_title_views.keys())
print(data_title_views)
```

```
{'How I Would Learn Data Science (If I Had to Start Over)': '1252970', 'The Best Free Data Science Courses Nobody is Talking
About': '297050', '3 Proven Data Science Projects for Beginners (Kaggle)': '237192', 'Beginner Kaggle Data Science Project Wa
lk-Through (Titanic)': '167881', 'The Projects You Should Do To Get A Data Science Job': '131573', 'How I Would Learn Data Sc
ience in 2021 (What Has Changed?)': '123484', "Why You Probably Won't Become a Data Scientist": '108043', 'Data Science Proje
ct from Scratch - Part 1 (Project Planning)': '102708', 'Why I Quit Data Science': '98758', '3 Reasons You Should NOT Become
a Data Scientist': '93282', 'Data Science Certificate vs Bootcamp vs Masters Degree': '92300', 'How I Learned Data Science':
'87146', 'How I Would Learn Data Science in 2022 (If I Had to Start Over)': '77307', 'How to Set Up Your Data Science Environ
ment (Anaconda Beginner)': '75207', 'Data Science Project from Scratch - Part 2 (Data Collection)': '71055', 'Is Data Science
Dying?': '69900', 'How to Make A Data Science Portfolio Website with Github Pages': '69004', 'Data Science Advice for College
Students': '62121', 'How to ULTRALEARN Data Science': '55294', 'What is the #66DaysOfData?': '52811', '5 Essential Data Scien
ce Projects for Your Portfolio': '51025', 'How YOU Can Land a Sports Analytics Job': '50447', 'Data Science Project from Scra
tch - Part 3 (Data Cleaning)': '50173', "Why I'm Starting Data Science Over Again.": '49559', 'Math Needed for Mastering Data
Science': '48363', 'The 7 Biggest Data Science  Beginner Mistakes': '48181', 'Data Science Project from Scratch - Part 4 (Exp
loratory Data Analysis)': '47138', 'Different Data Science Roles Explained (by a Data Scientist)': '44953', 'How to Build a D
ata Science Portfolio Website with Hugo & Github Pages [feat. Data Professor]': '44871', 'Is Data Science Right For You?': '4
4034', 'Scrape Twitter Data in Python with Twitterscraper Module': '41486', 'How to Go From Data Analyst to Data Scientist':
'40169', '9 Ways You Can Make Extra Income as a Data Scientist': '39327', 'The Data Science Projects that Got Me a Job': '333
77', 'Data Science Project from Scratch  - Part 5 (Model Building)': '27619', 'The Best Computer for Data Science Beginners':
'27566', 'Work From Home Data Scientist: Day in the Life': '26582', 'How To Get Data Science Experience (Without a Job)': '26
```

In [13]:
```
#turn the dic into two list: one for keys another for value. later use them for plot
title_list=list(data_title_views.keys())[0:15]
views_list=list(data_title_views.values())[0:15]
print(title_list)
print(views_list)
```

```
['How I Would Learn Data Science (If I Had to Start Over)', 'The Best Free Data Science Courses Nobody is Talking About', '3 Pr
oven Data Science Projects for Beginners (Kaggle)', 'Beginner Kaggle Data Science Project Walk-Through (Titanic)', 'The Project
s You Should Do To Get A Data Science Job', 'How I Would Learn Data Science in 2021 (What Has Changed?)', "Why You Probably Wo
n't Become a Data Scientist", 'Data Science Project from Scratch - Part 1 (Project Planning)', 'Why I Quit Data Science', '3 Re
asons You Should NOT Become a Data Scientist', 'Data Science Certificate vs Bootcamp vs Masters Degree', 'How I Learned Data Sc
ience', 'How I Would Learn Data Science in 2022 (If I Had to Start Over)', 'How to Set Up Your Data Science Environment (Anacon
da Beginner)', 'Data Science Project from Scratch - Part 2 (Data Collection)']
['1252970', '297050', '237192', '167881', '131573', '123484', '108043', '102708', '98758', '93282', '92300', '87146', '77307',
'75207', '71055']
```

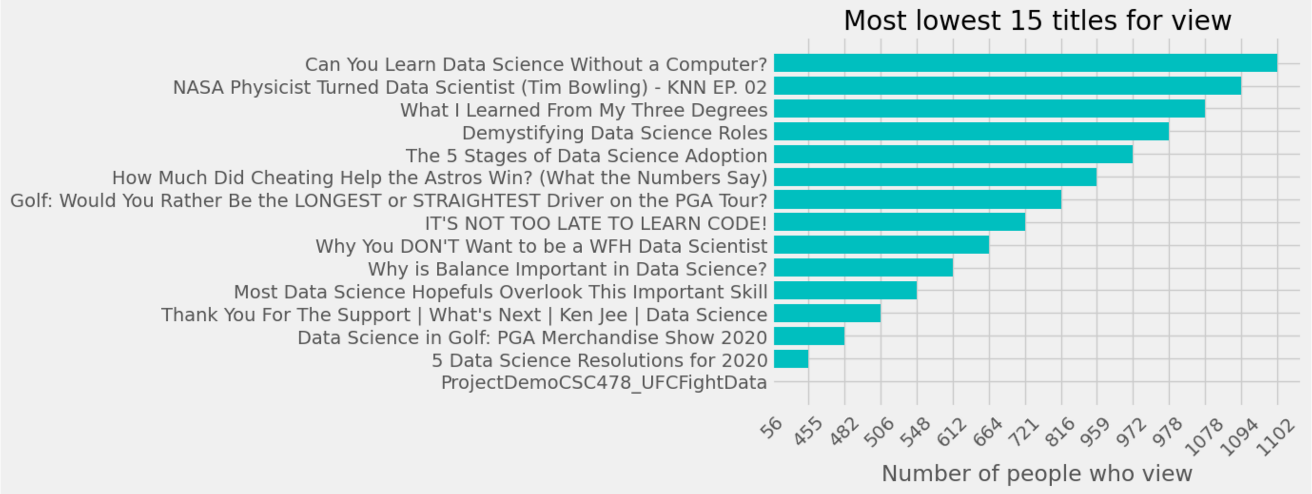
In [147]:
```
title_list.reverse()
views_list.reverse()
plt.barh(title_list,views_list)
plt.title('Most top 15 titles for view')
plt.xlabel('number of people who view')
plt.xticks(rotation=45)
plt.show()
```



```
In [146]: # get the freeze-15 titles people don't like to view
          title_list2=list(data_title_views.keys())[-15:]
          views_list2=list(data_title_views.values())[-15:]
          print(title_list2)
          print(views_list2)
```

```
['Can You Learn Data Science Without a Computer?', 'NASA Physicist Turned Data Scientist (Tim Bowling) - KNN EP. 02', 'What I L
earned From My Three Degrees', 'Demystifying Data Science Roles', 'The 5 Stages of Data Science Adoption', 'How Much Did Cheati
ng Help the Astros Win? (What the Numbers Say)', 'Golf: Would You Rather Be the LONGEST or STRAIGHTEST Driver on the PGA Tou
r?', "IT'S NOT TOO LATE TO LEARN CODE!", "Why You DON'T Want to be a WFH Data Scientist", 'Why is Balance Important in Data Sci
ence?', 'Most Data Science Hopefuls Overlook This Important Skill', "Thank You For The Support | What's Next | Ken Jee | Data S
cience", 'Data Science in Golf: PGA Merchandise Show 2020', '5 Data Science Resolutions for 2020', 'ProjectDemoCSC478_UFCFightD
ata']
['1102', '1094', '1078', '978', '972', '959', '816', '721', '664', '612', '548', '506', '482', '455', '56']
```

```
In [157]: title_list2.reverse()
          views_list2.reverse()
          plt.barh(title_list2,width=views_list2,color='c')
          plt.title('Most lowest 15 titles for view')
          plt.xlabel('Number of people who view')
          plt.xticks(rotation=45)
          plt.show()
```



```
In [4]: pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (3.7)Note: you may
need to restart the kernel to use updated packages.

Requirement already satisfied: joblib in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from nltk) (1.
1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from n
ltk) (2022.10.31)
Requirement already satisfied: click in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from nltk) (8.1.
3)
Requirement already satisfied: tqdm in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from nltk) (4.64.
1)
Requirement already satisfied: colorama in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from click->n
ltk) (0.4.5)
```

```
In [20]: pip install spacy
```

```
Collecting spacyNote: you may need to restart the kernel to use updated packages.

  Downloading spacy-3.4.3-cp310-cp310-win_amd64.whl (11.9 MB)
     -------------------------------------- 11.9/11.9 MB 179.6 kB/s eta 0:00:00
Requirement already satisfied: setuptools in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from spacy)
(63.2.0)
Requirement already satisfied: jinja2 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from spacy) (3.
1.2)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (fr
om spacy) (4.64.1)
Collecting pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4
  Downloading pydantic-1.10.2-cp310-cp310-win_amd64.whl (2.1 MB)
     ---------------------------------------- 2.1/2.1 MB 163.7 kB/s eta 0:00:00
Collecting catalogue<2.1.0,>=2.0.6
  Downloading catalogue-2.0.8-py3-none-any.whl (17 kB)
Collecting thinc<8.2.0,>=8.1.0
  Downloading thinc-8.1.5-cp310-cp310-win_amd64.whl (1.3 MB)
     ---------------------------------------- 1.3/1.3 MB 164.1 kB/s eta 0:00:00
Collecting spacy-loggers<2.0.0,>=1.0.0
  Downloading spacy_loggers-1.0.3-py3-none-any.whl (9.3 kB)
Collecting wasabi<1.1.0,>=0.9.1
  Downloading wasabi-0.10.1-py3-none-any.whl (26 kB)
Requirement already satisfied: packaging>=20.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from s
pacy) (21.3)
Collecting preshed<3.1.0,>=3.0.2
  Downloading preshed-3.0.8-cp310-cp310-win_amd64.whl (94 kB)
     --------------------------------------- 94.7/94.7 kB 90.2 kB/s eta 0:00:00
Collecting typer<0.8.0,>=0.3.0
  Downloading typer-0.7.0-py3-none-any.whl (38 kB)
Requirement already satisfied: numpy>=1.15.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from spa
cy) (1.23.2)
Collecting langcodes<4.0.0,>=3.2.0
  Downloading langcodes-3.3.0-py3-none-any.whl (181 kB)
     ------------------------------------ 181.6/181.6 kB 104.5 kB/s eta 0:00:00
Collecting cymem<2.1.0,>=2.0.2
  Downloading cymem-2.0.7-cp310-cp310-win_amd64.whl (29 kB)
Collecting spacy-legacy<3.1.0,>=3.0.10
  Downloading spacy_legacy-3.0.10-py2.py3-none-any.whl (21 kB)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages
(from spacy) (2.28.1)
Collecting srsly<3.0.0,>=2.4.3
  Downloading srsly-2.4.5-cp310-cp310-win_amd64.whl (479 kB)
     ------------------------------------ 479.4/479.4 kB 100.1 kB/s eta 0:00:00
Collecting murmurhash<1.1.0,>=0.28.0
  Downloading murmurhash-1.0.9-cp310-cp310-win_amd64.whl (18 kB)
Collecting pathy>=0.3.5
  Downloading pathy-0.10.0-py3-none-any.whl (48 kB)
     -------------------------------------- 48.9/48.9 kB 246.0 kB/s eta 0:00:00
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\yy\appdata\local\programs\python\python310\lib\site-package
s (from packaging>=20.0->spacy) (3.0.9)
Collecting smart-open<6.0.0,>=5.2.1
  Downloading smart_open-5.2.1-py3-none-any.whl (58 kB)
     --------------------------------------- 58.6/58.6 kB 52.5 kB/s eta 0:00:00
Collecting typing-extensions>=4.1.0
  Downloading typing_extensions-4.4.0-py3-none-any.whl (26 kB)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\yy\appdata\local\programs\python\python310\lib\site-package
s (from requests<3.0.0,>=2.13.0->spacy) (2.1.1)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (fro
m requests<3.0.0,>=2.13.0->spacy) (2022.9.24)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages
(from requests<3.0.0,>=2.13.0->spacy) (1.26.13)
Requirement already satisfied: idna<4,>=2.5 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from requ
ests<3.0.0,>=2.13.0->spacy) (3.4)
Collecting blis<0.8.0,>=0.7.8
  Downloading blis-0.7.9-cp310-cp310-win_amd64.whl (7.0 MB)
     ---------------------------------------- 7.0/7.0 MB 98.8 kB/s eta 0:00:00
Collecting confection<1.0.0,>=0.0.1
  Downloading confection-0.0.3-py3-none-any.whl (32 kB)
Requirement already satisfied: colorama in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from tqdm<5.
0.0,>=4.38.0->spacy) (0.4.5)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (fr
om typer<0.8.0,>=0.3.0->spacy) (8.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from j
inja2->spacy) (2.1.1)
Installing collected packages: wasabi, cymem, typing-extensions, spacy-loggers, spacy-legacy, smart-open, murmurhash, langcode
s, catalogue, blis, typer, srsly, pydantic, preshed, pathy, confection, thinc, spacy
Successfully installed blis-0.7.9 catalogue-2.0.8 confection-0.0.3 cymem-2.0.7 langcodes-3.3.0 murmurhash-1.0.9 pathy-0.10.0 pr
eshed-3.0.8 pydantic-1.10.2 smart-open-5.2.1 spacy-3.4.3 spacy-legacy-3.0.10 spacy-loggers-1.0.3 srsly-2.4.5 thinc-8.1.5 typer-
0.7.0 typing-extensions-4.4.0 wasabi-0.10.1
```

In [33]:

```
pip install gensim
```

```
Collecting gensim
  Downloading gensim-4.2.0-cp310-cp310-win_amd64.whl (23.9 MB)
     --------------------------------------- 23.9/23.9 MB 36.8 kB/s eta 0:00:00
Requirement already satisfied: numpy>=1.17.0 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from gen
sim) (1.23.2)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from
gensim) (5.2.1)
Requirement already satisfied: scipy>=0.18.1 in c:\users\yy\appdata\local\programs\python\python310\lib\site-packages (from gen
sim) (1.9.1)
Collecting Cython==0.29.28
  Downloading Cython-0.29.28-py2.py3-none-any.whl (983 kB)
     ------------------------------------- 983.8/983.8 kB 63.3 kB/s eta 0:00:00
Installing collected packages: Cython, gensim
Successfully installed Cython-0.29.28 gensim-4.2.0
Note: you may need to restart the kernel to use updated packages.
```

In [48]:

```
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import GridSearchCV
from pprint import pprint

from spacy.cli import download
download("en_core_web_sm")
import re, nltk, spacy, gensim
import matplotlib.pyplot as plt
%matplotlib inline
```

```
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [29]:

```
import pandas as pd
#df = pd.read_csv('hotelreviews.csv', encoding='utf-8')
df = pd.read_csv('data_title_views.csv')
df.info()
df.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 223 entries, 0 to 222
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Title   223 non-null    object
 1   Views   223 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.6+ KB
```

Out[29]:

	Title	Views
0	How I Would Learn Data Science (If I Had to St...	1252970
1	The Best Free Data Science Courses Nobody is T...	297050
2	3 Proven Data Science Projects for Beginners (...	237192
3	Beginner Kaggle Data Science Project Walk-Thro...	167881
4	The Projects You Should Do To Get A Data Scien...	131573
5	How I Would Learn Data Science in 2021 (What H...	123484
6	Why You Probably Won't Become a Data Scientist	108043
7	Data Science Project from Scratch - Part 1 (Pr...	102708
8	Why I Quit Data Science	98758
9	3 Reasons You Should NOT Become a Data Scientist	93282

In [39]:

```
# df = pd.read_csv('data_title_views.csv')
def sent_to_words(sentences):
    for sent in sentences:
        sent = re.sub("\'", "", sent)  # remove single quotes
        sent = gensim.utils.simple_preprocess(str(sent), deacc=True)
        yield(sent)

# Convert to list
data_words = list(sent_to_words(list(data_title_views.keys())[0:50]))
print(data_words)
```

```
[['how', 'would', 'learn', 'data', 'science', 'if', 'had', 'to', 'start', 'over'], ['the', 'best', 'free', 'data', 'science',
'courses', 'nobody', 'is', 'talking', 'about'], ['proven', 'data', 'science', 'projects', 'for', 'beginners', 'kaggle'], ['begi
nner', 'kaggle', 'data', 'science', 'project', 'walk', 'through', 'titanic'], ['the', 'projects', 'you', 'should', 'do', 'to',
'get', 'data', 'science', 'job'], ['how', 'would', 'learn', 'data', 'science', 'in', 'what', 'has', 'changed'], ['why', 'you',
'probably', 'wont', 'become', 'data', 'scientist'], ['data', 'science', 'project', 'from', 'scratch', 'part', 'project', 'plann
ing'], ['why', 'quit', 'data', 'science'], ['reasons', 'you', 'should', 'not', 'become', 'data', 'scientist'], ['data', 'scienc
e', 'certificate', 'vs', 'bootcamp', 'vs', 'masters', 'degree'], ['how', 'learned', 'data', 'science'], ['how', 'would', 'lear
n', 'data', 'science', 'in', 'if', 'had', 'to', 'start', 'over'], ['how', 'to', 'set', 'up', 'your', 'data', 'science', 'enviro
nment', 'anaconda', 'beginner'], ['data', 'science', 'project', 'from', 'scratch', 'part', 'data', 'collection'], ['is', 'dat
a', 'science', 'dying'], ['how', 'to', 'make', 'data', 'science', 'portfolio', 'website', 'with', 'github', 'pages'], ['data',
'science', 'advice', 'for', 'college', 'students'], ['how', 'to', 'ultralearn', 'data', 'science'], ['what', 'is', 'the', 'days
ofdata'], ['essential', 'data', 'science', 'projects', 'for', 'your', 'portfolio'], ['how', 'you', 'can', 'land', 'sports', 'an
alytics', 'job'], ['data', 'science', 'project', 'from', 'scratch', 'part', 'data', 'cleaning'], ['why', 'im', 'starting', 'dat
a', 'science', 'over', 'again'], ['math', 'needed', 'for', 'mastering', 'data', 'science'], ['the', 'biggest', 'data', 'scienc
e', 'beginner', 'mistakes'], ['data', 'science', 'project', 'from', 'scratch', 'part', 'exploratory', 'data', 'analysis'], ['di
fferent', 'data', 'science', 'roles', 'explained', 'by', 'data', 'scientist'], ['how', 'to', 'build', 'data', 'science', 'portf
olio', 'website', 'with', 'hugo', 'github', 'pages', 'feat', 'data', 'professor'], ['is', 'data', 'science', 'right', 'for', 'y
ou'], ['scrape', 'twitter', 'data', 'in', 'python', 'with', 'twitterscraper', 'module'], ['how', 'to', 'go', 'from', 'data', 'a
nalyst', 'to', 'data', 'scientist'], ['ways', 'you', 'can', 'make', 'extra', 'income', 'as', 'data', 'scientist'], ['the', 'dat
a', 'science', 'projects', 'that', 'got', 'me', 'job'], ['data', 'science', 'project', 'from', 'scratch', 'part', 'model', 'bui
lding'], ['the', 'best', 'computer', 'for', 'data', 'science', 'beginners'], ['work', 'from', 'home', 'data', 'scientist', 'da
y', 'in', 'the', 'life'], ['how', 'to', 'get', 'data', 'science', 'experience', 'without', 'job'], ['kaggle', 'project', 'fro
m', 'scratch', 'part', 'data', 'science', 'profession', 'survey'], ['how', 'to', 'learn', 'programming', 'for', 'data', 'scienc
e', 'steps'], ['what', 'does', 'data', 'scientist', 'actually', 'do'], ['how', 'subscriber', 'landed', 'data', 'analyst', 'jo
b', 'in', 'less', 'than', 'year', 'ray', 'ojel', 'knn', 'ep'], ['wish', 'had', 'known', 'this', 'before', 'starting', 'in', 'da
ta', 'science'], ['how', 'learned', 'to', 'learn'], ['avoid', 'these', 'data', 'science', 'resume', 'mistakes'], ['uber', 'driv
er', 'to', 'machine', 'learning', 'engineer', 'in', 'months', 'daniel', 'bourke', 'knn', 'ep'], ['reviewing', 'your', 'data',
'science', 'resumes', 'episode', 'different', 'resumes'], ['built', 'the', 'first', 'ever', 'youtube', 'subscriber', 'leaderboa
rd'], ['reviewing', 'your', 'data', 'science', 'projects', 'episode', 'very', 'detailed', 'project'], ['how', 'got', 'my', 'fir
st', 'data', 'science', 'internship', 'and', 'how', 'you', 'can', 'land', 'one']]
```

In [69]:

```
def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV','WH','FW']): #'NOUN', 'ADJ', 'VERB', 'ADV'
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append(" ".join([token.lemma_ if token.lemma_ not in ['-PRON-'] else '' for token in doc if token.pos_ in allow
    return texts_out
```

In [70]:

```
# spacy for lemmatization
# Initialize spacy ‘en’ model, keeping only tagger component (for efficiency)
# Run in terminal: python -m spacy download en

nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])

# Do lemmatization keeping only Noun, Adj, Verb, Adverb, WH, FW
data_lemmatized = lemmatization(data_words, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV','WH','FW'])

print(data_lemmatized)
```

```
['learn data science start', 'well free datum science course talk', 'prove data science project beginner kaggle', 'beginner sci
ence project walk titanic', 'project get datum science job', 'learn data science change', 'probably become data scientist', 'da
ta science project scratch part project planning', 'quit datum science', 'reason become data scientist', 'data science certific
ate vs bootcamp master degree', 'learn data science', 'learn data science start', 'set data science environment', 'data science
project scratch part datum collection', 'data science die', 'make data science portfolio website page', 'datum science advice c
ollege student', 'ultralearn data science', 'daysofdata', 'essential datum science project portfolio', 'land sport analytic jo
b', 'data science project scratch part datum cleaning', 'm start datum science over again', 'math need master data science', 'b
ig data science beginner mistake', 'data science project scratch part exploratory datum analysis', 'different datum science rol
e explain datum scientist', 'build datum science portfolio website page feat data professor', 'data science right', 'scrape twi
tter datum python twitterscraper module', 'go datum analyst data scientist', 'way make extra income data scientist', 'data scie
nce project get job', 'data science project scratch part model building', 'good computer data science beginner', 'work home dat
um day life', 'get data science experience job', 'project scratch part science profession survey', 'learn program datum science
step', 'data scientist actually', 'land analyst job less year ray', 'wish know start data science', 'learn learn', 'avoid datum
science resume mistake', 'driver machine learning engineer month knn', 'review data science resume episode different resume',
'build first ever youtube subscriber leaderboard', 'review data science project episode very detailed project', 'get first data
science internship land']
```

In [51]:
```
vectorizer = CountVectorizer(analyzer='word',
                             min_df=3,                        # minimum reqd occurences of a word
                             stop_words='english',             # remove stop words
                             lowercase=True,                   # convert all words to lowercase
                             token_pattern='[a-zA-Z0-9]{3,}',  # num chars > 3
                             # max_features=50000,             # max number of uniq words
                            )

data_vectorized = vectorizer.fit_transform(data_lemmatized)
```

In [52]:
```
vectorizer
```

Out[52]:
```
CountVectorizer(min_df=3, stop_words='english', token_pattern='[a-zA-Z0-9]{3,}')
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

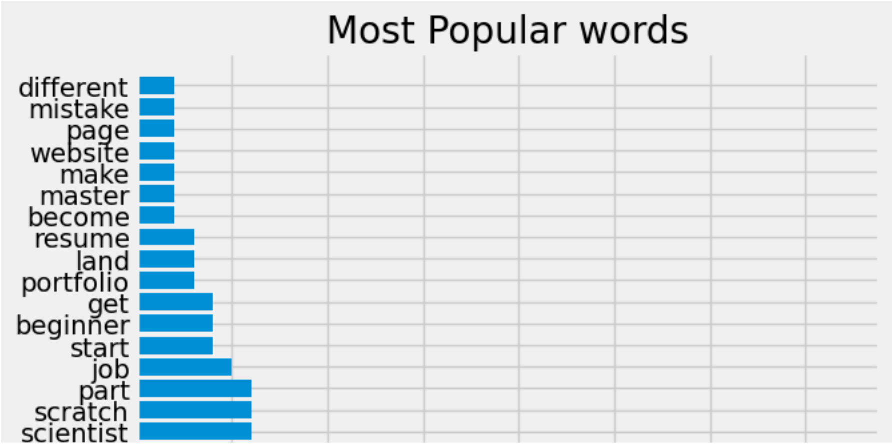
In [55]:
```
#print(data_vectorized)
```

In [64]:
```

```

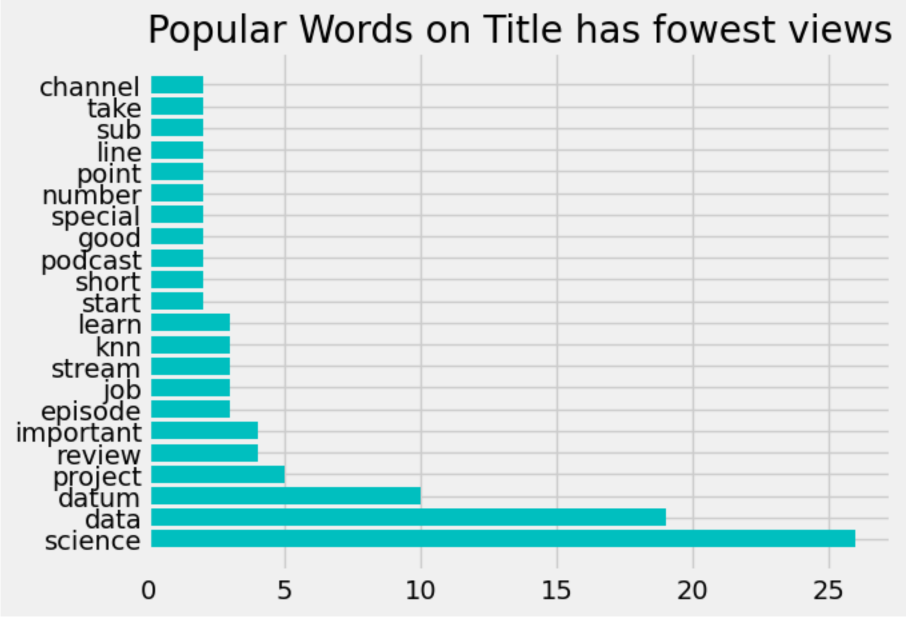
In [72]:
```
# count the frequency of words
from collections import Counter
words_counter=Counter()
for row in data_lemmatized:
    words_counter.update(row.split(' '))
print(words_counter)
```

```
Counter({'science': 37, 'data': 31, 'datum': 17, 'project': 14, 'learn': 7, 'scientist': 6, 'scratch': 6, 'part': 6, 'job': 5,
'start': 4, 'beginner': 4, 'get': 4, 'portfolio': 3, 'land': 3, 'resume': 3, 'become': 2, 'master': 2, 'make': 2, 'website': 2,
'page': 2, 'mistake': 2, 'different': 2, 'build': 2, 'analyst': 2, 'review': 2, 'episode': 2, 'first': 2, 'well': 1, 'free': 1,
'course': 1, 'talk': 1, 'prove': 1, 'kaggle': 1, 'walk': 1, 'titanic': 1, 'change': 1, 'probably': 1, 'planning': 1, 'quit': 1,
'reason': 1, 'certificate': 1, 'vs': 1, 'bootcamp': 1, 'degree': 1, 'set': 1, 'environment': 1, 'collection': 1, 'die': 1, 'adv
ice': 1, 'college': 1, 'student': 1, 'ultralearn': 1, 'daysofdata': 1, 'essential': 1, 'sport': 1, 'analytic': 1, 'cleaning':
1, 'm': 1, 'over': 1, 'again': 1, 'math': 1, 'need': 1, 'big': 1, 'exploratory': 1, 'analysis': 1, 'role': 1, 'explain': 1, 'fe
at': 1, 'professor': 1, 'right': 1, 'scrape': 1, 'twitter': 1, 'python': 1, 'twitterscraper': 1, 'module': 1, 'go': 1, 'way':
1, 'extra': 1, 'income': 1, 'model': 1, 'building': 1, 'good': 1, 'computer': 1, 'work': 1, 'home': 1, 'day': 1, 'life': 1, 'ex
perience': 1, 'profession': 1, 'survey': 1, 'program': 1, 'step': 1, 'actually': 1, 'less': 1, 'year': 1, 'ray': 1, 'wish': 1,
'know': 1, 'avoid': 1, 'driver': 1, 'machine': 1, 'learning': 1, 'engineer': 1, 'month': 1, 'knn': 1, 'ever': 1, 'youtube': 1,
'subscriber': 1, 'leaderboard': 1, 'very': 1, 'detailed': 1, 'internship': 1})
```

In [80]:
```
words=[]
popularity=[]
for item in words_counter.most_common(22):
    words.append(item[0])
    popularity.append(item[1])
plt.style.use('fivethirtyeight')
plt.barh(words, popularity)
plt.title('Most Popular words')
#plt.xticks(rotation=45)
plt.show()
```



In [90]:
```
data_words2 = list(sent_to_words(list(data_title_views.keys())[-50:]))
#print(data_words2)

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV','WH','FW']): #'NOUN', 'ADJ', 'VERB', 'ADV'
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append(" ".join([token.lemma_ if token.lemma_ not in ['-PRON-'] else '' for token in doc if token.pos_ in allow
    return texts_out

nlp2 = spacy.load('en_core_web_sm', disable=['parser', 'ner'])
data_lemmatized2 = lemmatization(data_words2, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV','WH','FW'])
#print(data_lemmatized2)

from collections import Counter
words_counter2=Counter()
for row2 in data_lemmatized2:
    words_counter2.update(row2.split(' '))
#print(words_counter2)

words2=[]
popularity2=[]
for item in words_counter2.most_common(22):
    words2.append(item[0])
    popularity2.append(item[1])

plt.style.use('fivethirtyeight')
plt.barh(words2, popularity2, color='c')
plt.title('Popular Words on Title has fowest views')
#plt.xticks(rotation=45)
plt.show()
```



In [ ]:

In [ ]:
```
# Modeling at Weka App
```

In [ ]:

In [ ]:

In [ ]: