



SI206 Final Project
MVE
Elaina Tiller
Victoria Stahl
Madeleine Singh





GOALS

Initial Goals

- APIs used: Twitter, iTunes, Spotify
- Use the frequency of keyword/podcast name on Twitter and compare it with the top podcast charts of iTunes and Spotify to see if they represented the public's favorite podcasts.

Final Goals

- APIs/websites used: YouTube, Spotify, Reddit
- Compare data on The Joe Rogan Experience Podcast such as release date, popularity, and guest appearances to gauge popularity of guests on the 1600 episodes available and popularity of the reddit page




?

Problems Faced



Problem


All videos were removed from Youtube due to Rogan's exclusive Spotify deal. Our Youtube API gave the most data for our database and we didn't want to lose it all. Changed code to pull data from a csv and only access videos before a week ago.




Problem : two tables for 1 API with guest names

**Titles of the videos
were irregular**


**Could only put one guestid
to JOIN the tables, so invalid
id for tables with 2 guests (or
clips) is 0**



Joe Rogan Experience #1529 -
Whitney Cummings & Annie
Lederman
Joe Rogan Experience #1531 -
Miley Cyrus



**Used Regex to get
names of episodes
with one or two
people**

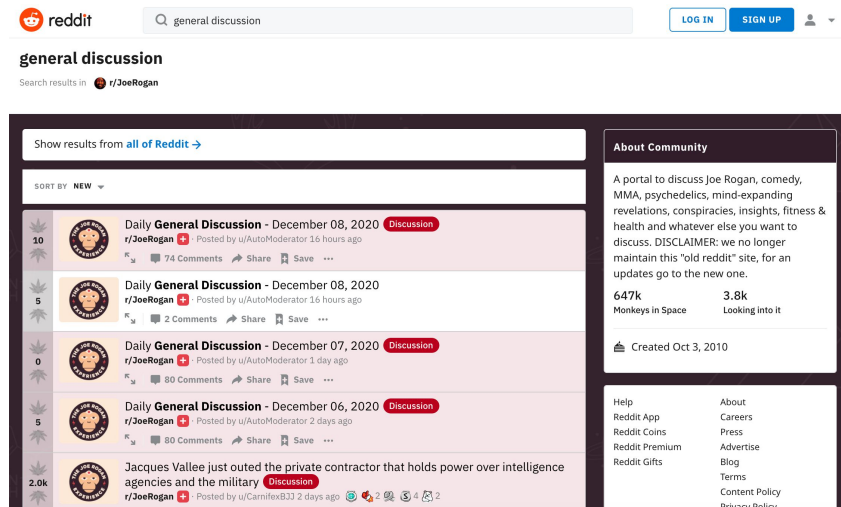


**When printing guest
names of episodes of
two people, couldn't
use guest id, had to use
title and regex**

Problem 3:

When scrapping Reddit for the discussion posts for Joe Rogan's podcast with the most comments by date, there were hundreds of discussion dates available once the webpage was loaded, but when scraping using .requests, it would only return the 17 most recent dates.

- Solution: In office hours we were recommended to download the webpage at a .htm file like in Project 2 and we then limited the data grabbed to 100 entries.



Calculations★



Guest,Number Apperances

Brendan Schaub,8

Tim Dillon,6

Tom Papa,6

Joey Diaz,6

Donnell Rawlings,5

Tony Hinchcliffe,5

Mike Baker,5

Brian Redban,5

Lex Fridman,4

Duncan Trussell,4

Greg Fitzsimmons,4

Andrew Santino,4

Steven Rinella,3

Bridget Phetasy,3

Jeremy Corbell,3

Bill Burr,3

Andrew Schulz,3

Michael Yo,3

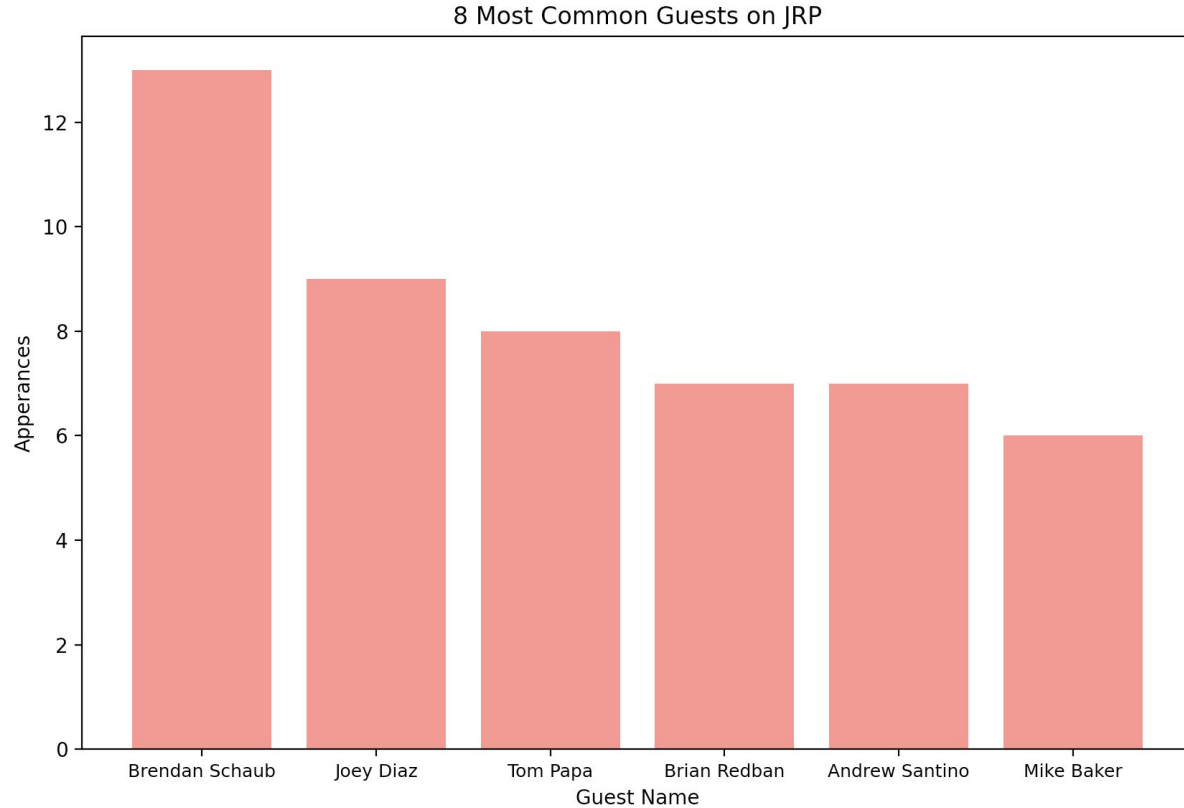
Tim Pool,3

Eric Weinstein,3

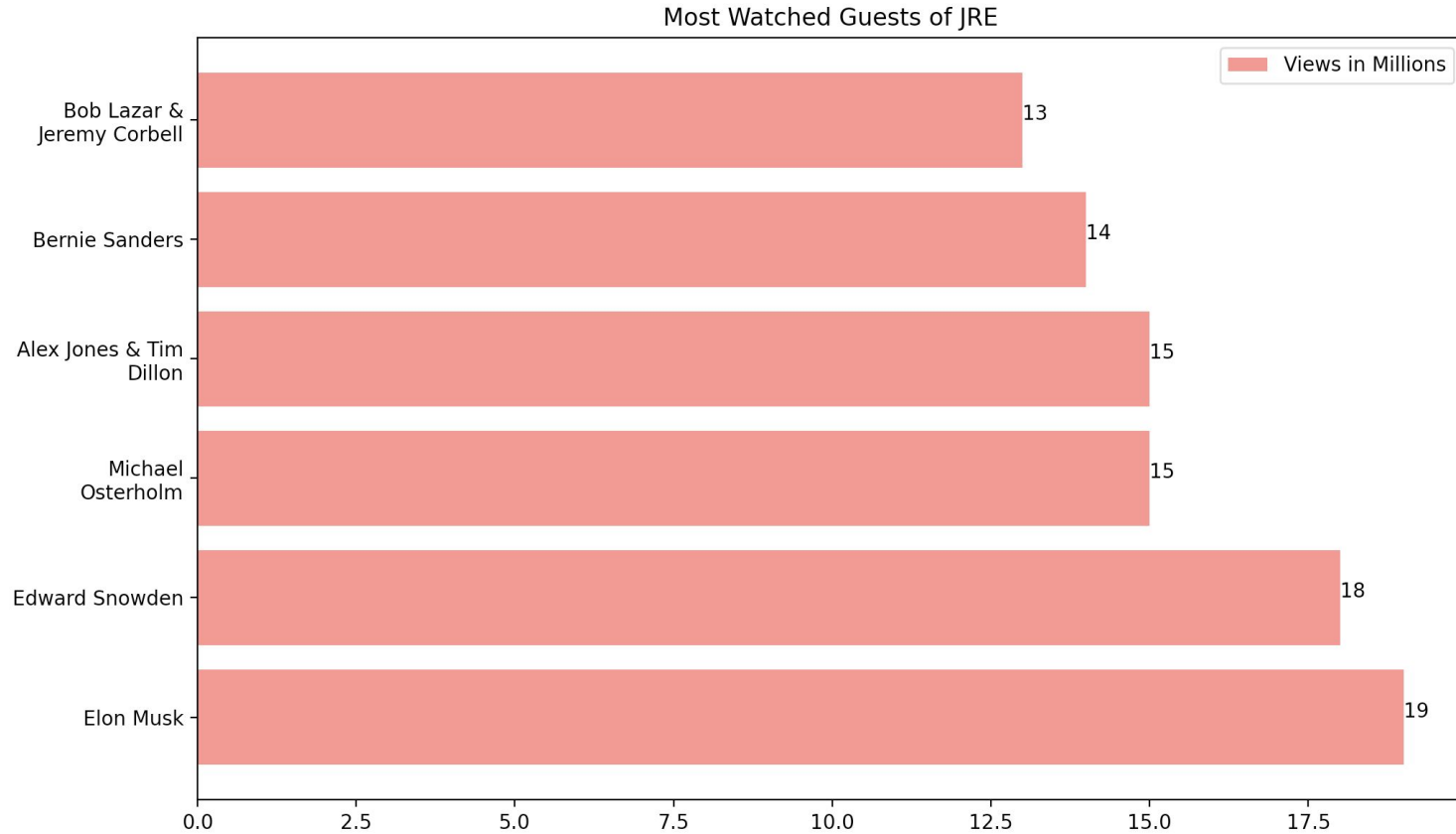
Michael Malice,3

1. The average number of comments on the 100 most recent Joe Rogan Experience discussion posts on Reddit is 90.69 comments.
2. Calculate guest appearances on shows in database

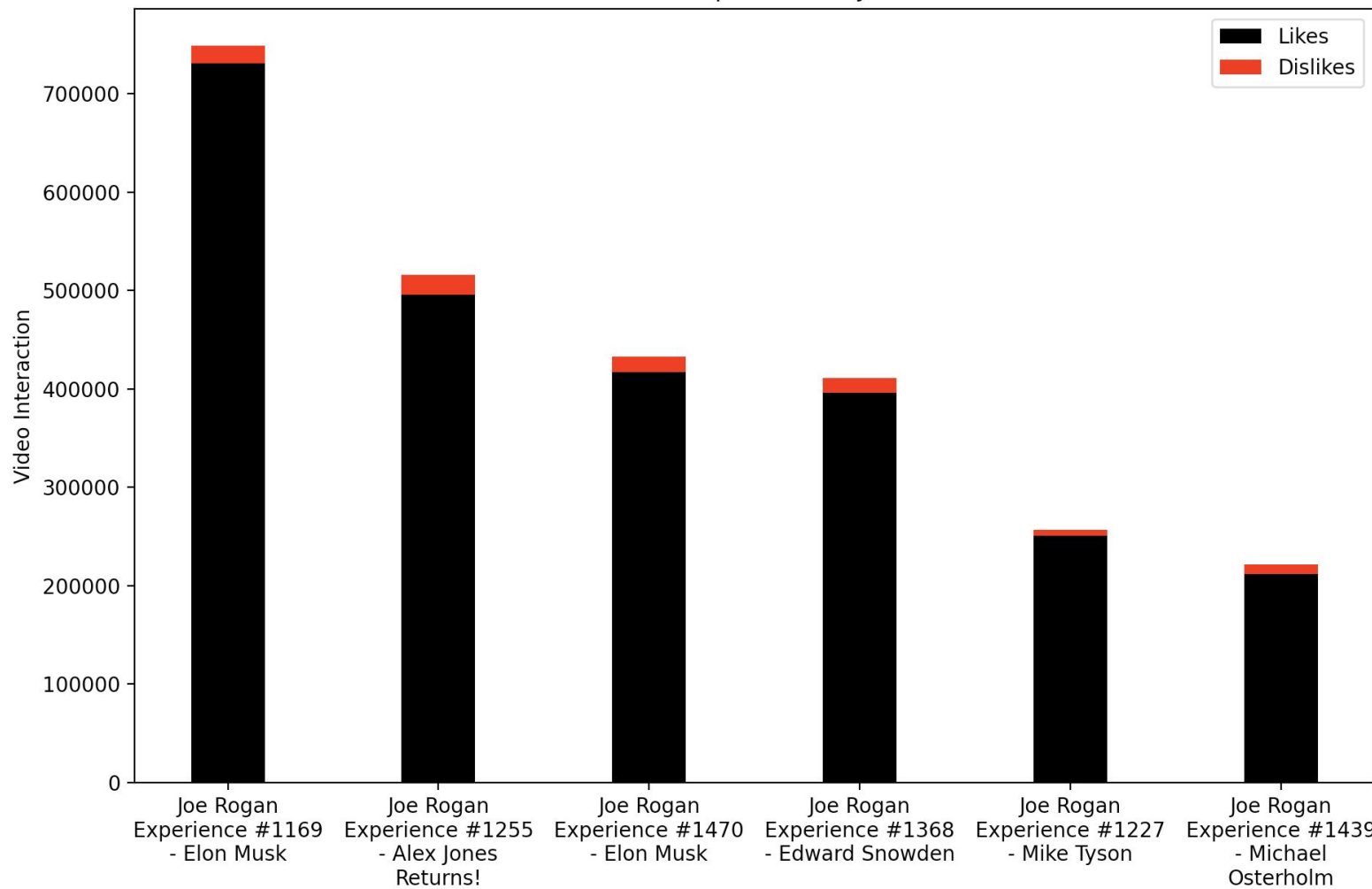
Visualization



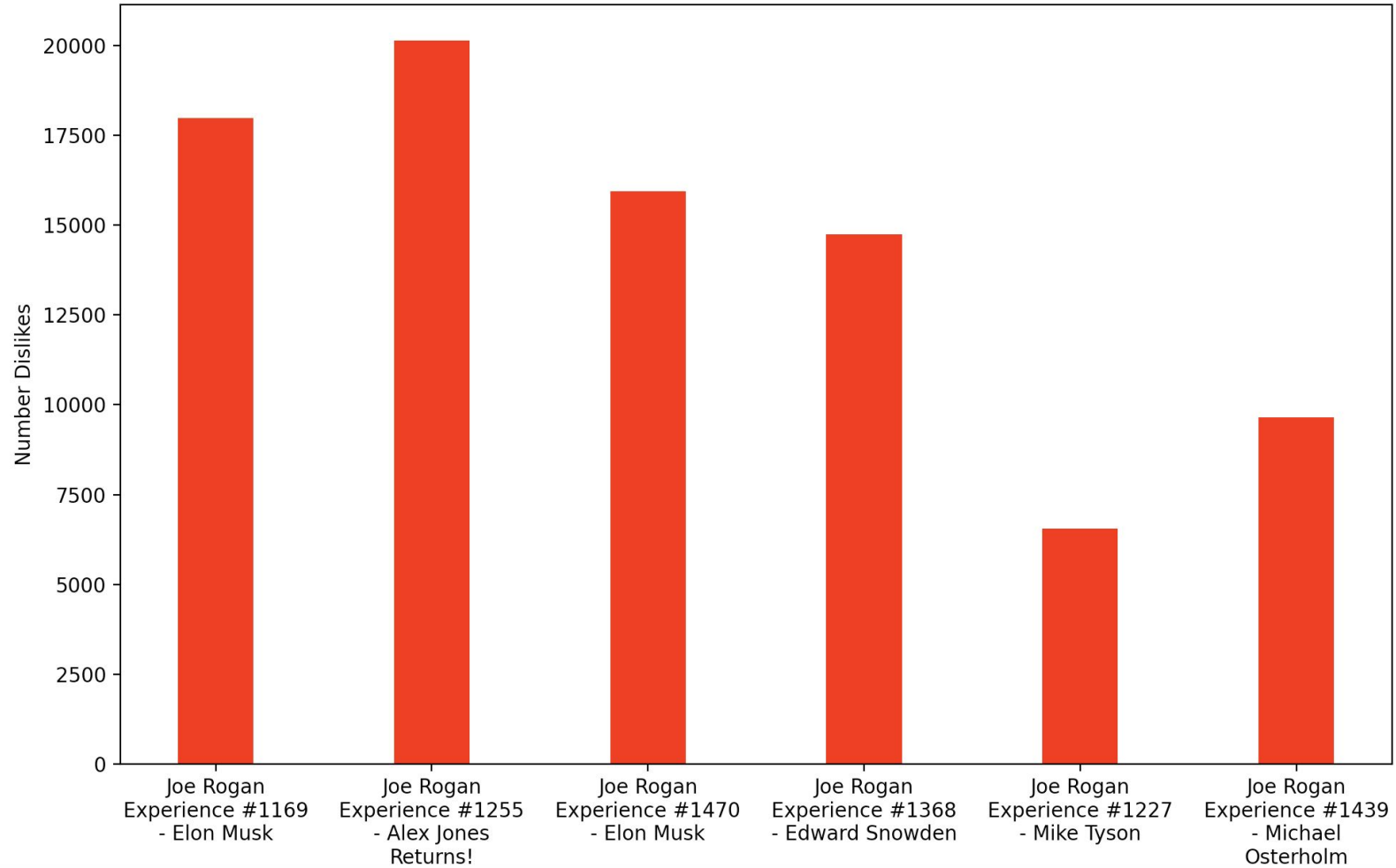
Visualization



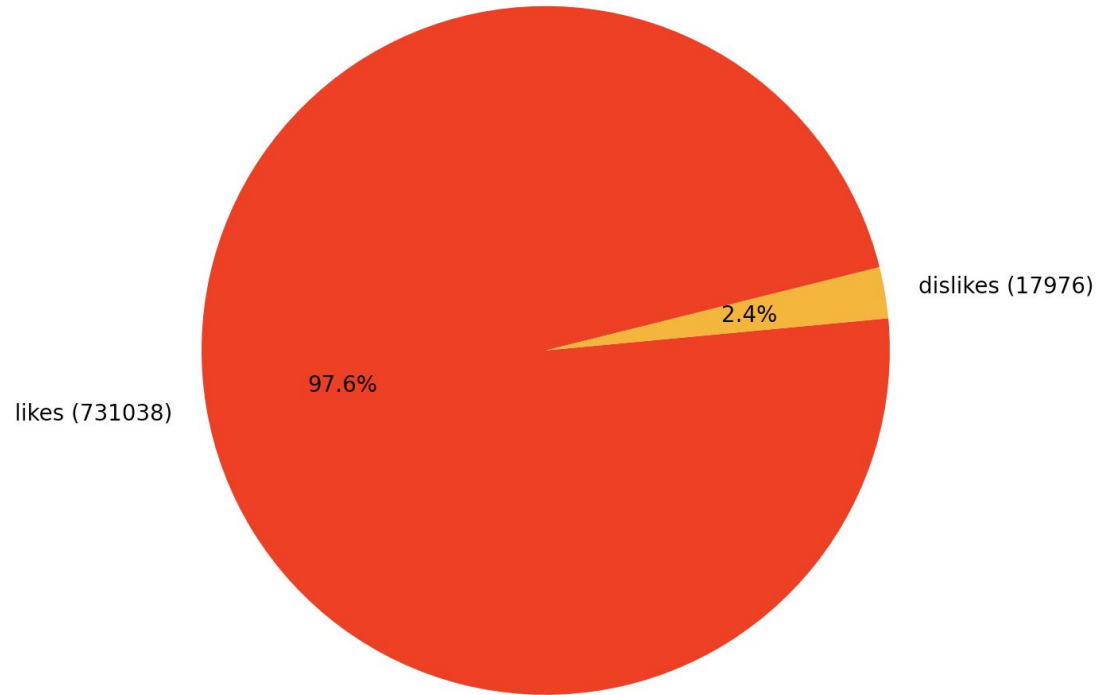
Video Interactions for Top 6 Viewed JRP Youtube videos



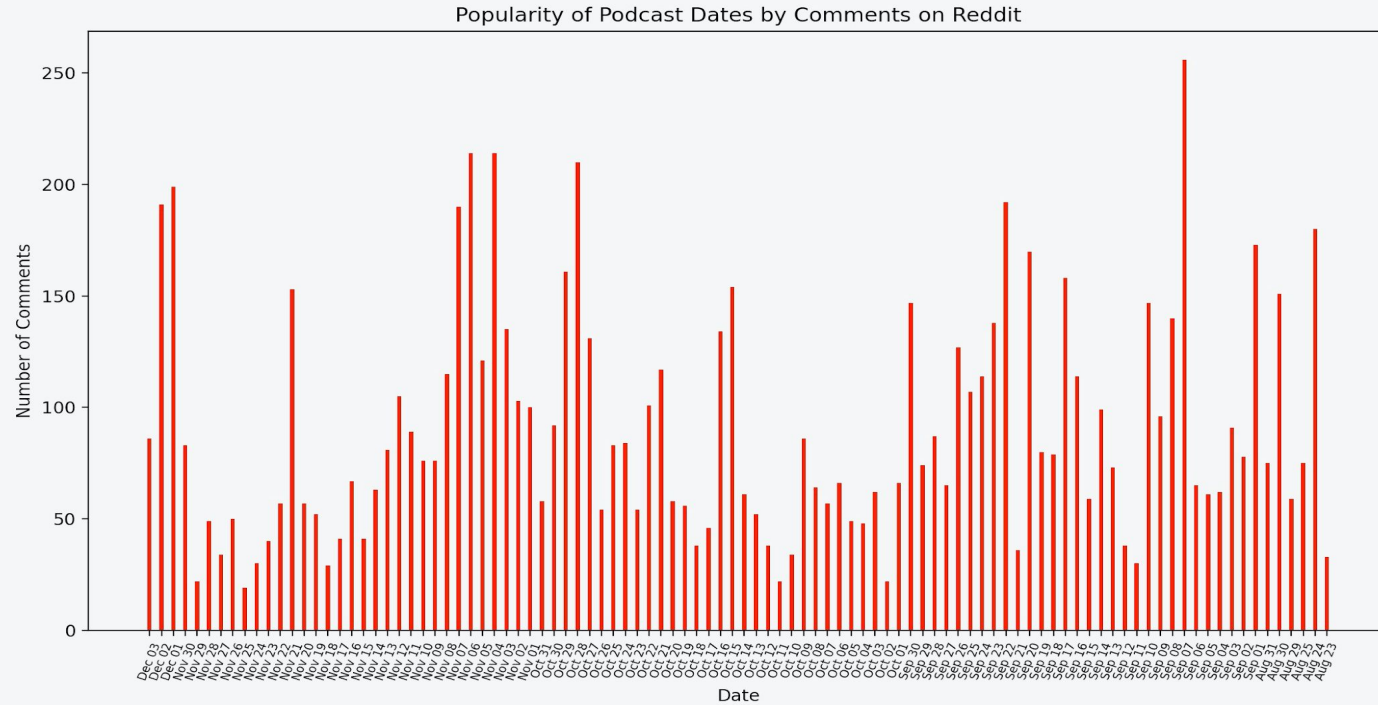
Top 6 Disliked Videos on JRP Youtube



Most Viewed Episode Joe Rogan Experience #1169 - Elon Musk likes to dislikes



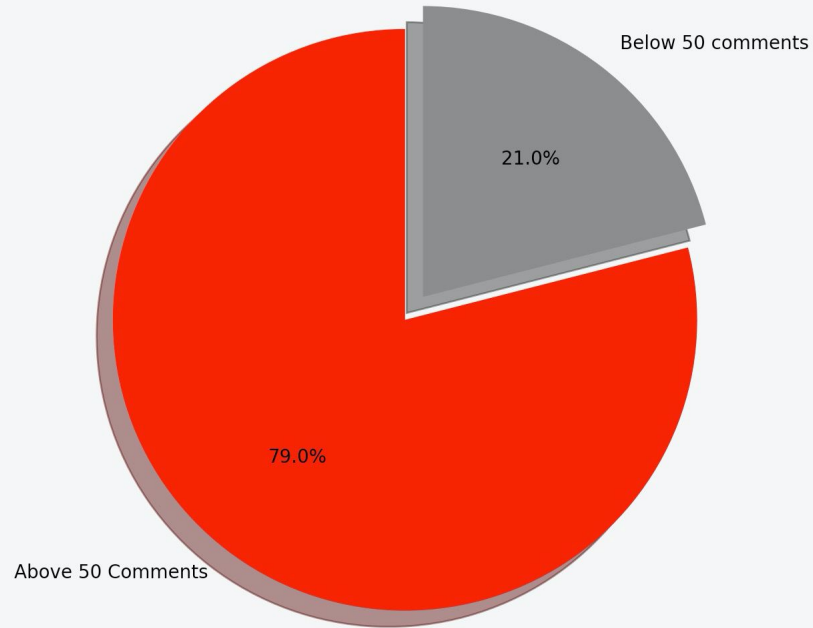
Visualization



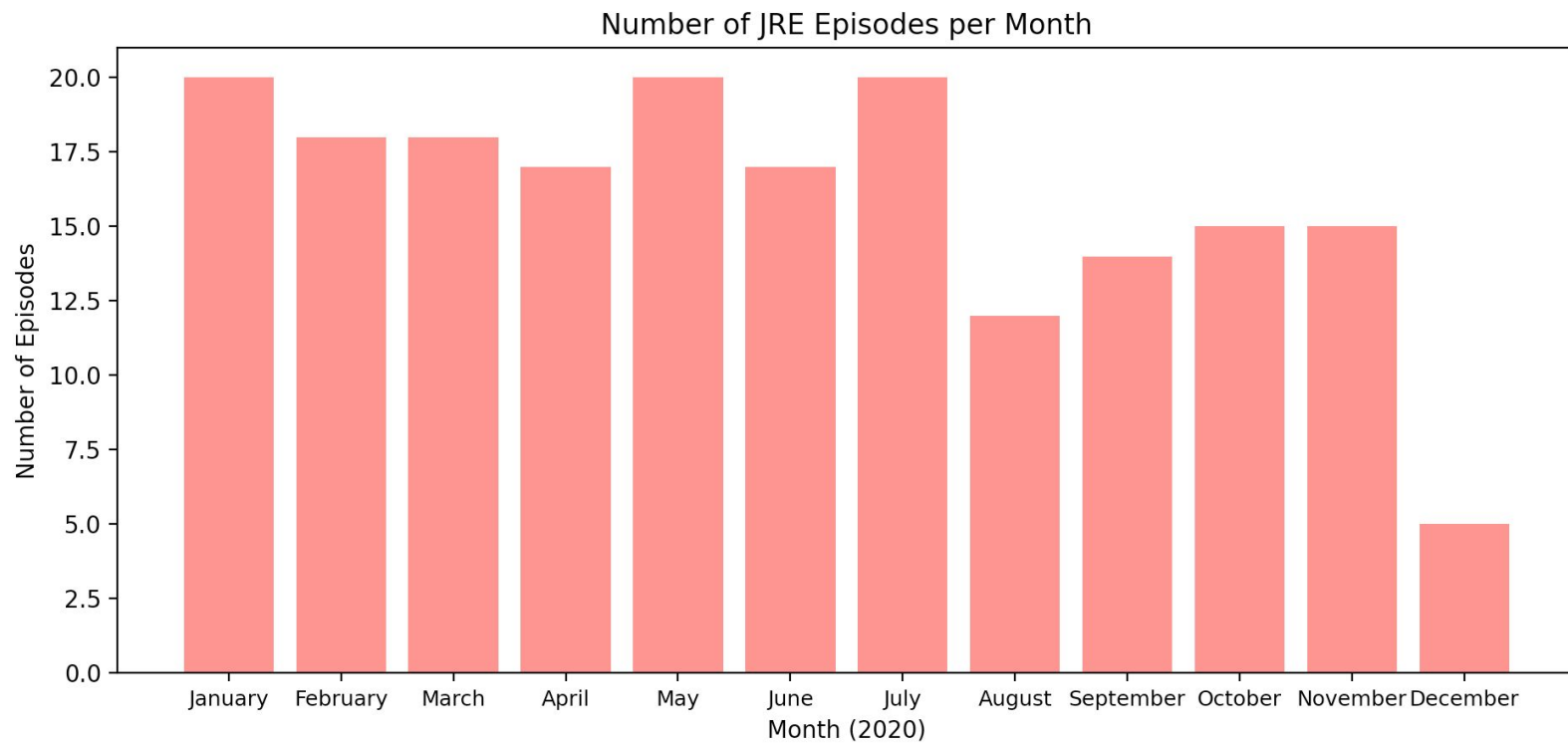
Visualization



Percentage of 100 Most Recent Joe Rogan Experience Discussion Posts on Reddit Above 50 Comments



Visualization



Visualization



Proportion of "Special Episodes" (episodes that are not numbered)



Documentation - Youtube_JRP.py

```
def readDataFromFile(filename):
    '''Takes in the file youtube_data.csv to read it and return file data'''

def setUpDatabase(db_name):
    '''Takes in the database 'JRP.bd' as a paramater and sets it up, also returning cur and conn'''

def uploadDataJRE(cur,conn):
    '''Takes in cur and conn to create JRP table in JRP.db with all the information for
    each podcasts on Rogans youtube channel. Finds the last inserted id, and inserts another 25
    more podcasts from the csv file that has not been added'''

def getNames(cur):
    '''Takes in the cursur to select all titles of podcasts id JRP(format is Joe Rogan Experience #1560 - Mike Baker)
    and use regex to get the guest name (Mike Baker) ONLY for podcasts of one or two people, otherwise it
    is an irregular post on the channel. Returns the list of names'''

def countNames(names):
    '''takes in list of names, and returns sorted dictionary of names with the number of apperances'''

def printNamesPretty(counts,file):
    '''Takes in a dictionary in form {guestnames:apperances} and writes out the data to youtube.txt file
    as a calulation'''

def putNamesInData(counts,cur,conn):
    '''Takes in the counts dictionary in the form {guestnames:apperances}, cur, and conn.
    Makes JRP_guest_count table in JRP.db with an id for each guest, their name, and
    number of apperances. The id serves to link two tables together. '''

def fillGuestId(cur,conn):
    '''Takes in cur and conn to insert guestid into JRP from JRP_guest_count. This common key
    allows us to JOIN tables and not have repeated guest names. Episodes titles that are irregular
    or have two guests get a guestid of 0'''
```

```
def barChart2(cur):  
    ''' Shows number of views in millions for the top 6 watched videos in the database'''  
  
def barChart3(cur):  
    '''Show number of likes and disliked stacked on top of eachother for top 6 most viewed videos  
    in the database'''  
  
def barChart4(cur):  
    '''Makes a bar graph of the Top 6 disliked videos in database and  
    their number of dislikes with episode title'''  
  
def pieChartMostViewedEps(cur):  
    '''Makes a pie chart of the most viewed episode in the database  
    with percentage and number of likes to dislikes displayed'''  
  
def main():  
    ''' calls everything above in sections'''
```

Documentation – reddit.py

```
def getDates(filename):
    '''This function takes the .htm file as a parameter and scrapes the file for the date of the discussion post
    and the number of comments on that discussion post. It returns a list of 100 tuples that include the date,
    number of comments, and count which we use later as discussion_id'''

def setUpDatabase(db_name):
    '''This function takes the database 'JRP.db' as a parameter, sets up the database, and returns cur and conn.'''

def setUpComments(dates_comments, cur, conn):
    '''This function takes the list of tuples dates_comments, cur, and conn as parameters and inserts the data from
    dates_comments into the database set up above (JRP.db). This code needs to be run 4 times since it inputs data
    25 rows at a time and has a total of 100 rows.'''

def getAverageComments(cur):
    '''This function selects the dates from the Popularity table to calculate the total number of dates and the comments
    from the Popularity table to calculate the total number of comments for the 100 dates. It then returns the average
    number of comments for each discussion post.'''

def printAverageComments(comments, file):
    '''This function writes the average number of comments to reddit.txt file as a calculation.'''

def makeVisualizations(cur):
    '''This function selects the dates and number of comments from the Popularity table and returns a bar chart
    with the number of comments on each of the 100 most recent Joe Rogan Experience discussion posts on Reddit.'''

def vizualizationByComments(cur):
    '''This function selects the dates from the Popularity table to calculate the total number of dates gathered
    and the number of comments from the Popularity table that are above 50 to make a percentage of discussion
    posts that have above 50 comments since those posts are considered 'popular'. '''

def main():
    '''This function calls the above functions.'''
```

Documentation – *spotify.py*

```
def episodes_search(id, offset, cur):
    '''this function uses the Spotify API to grab information for 25 episodes at a time (starting with the most
    recent, but can be offset by updating the offset parameter), and then finds the name and release date of each
    episode grabbed. it returns a list of 25 tuples of (name, release date)'''

def setUpDatabase(db_name):
    '''this function takes the database 'JRP.db' as a parameter, sets up the database, and returns cur and conn'''

def setUpEpisodes(data, cur, conn):
    '''this function sets up the Episodes table that will go into the JRP.db database. It takes the list of tuples
    returned by episodes_search and put them in a table that has values episode_id (which is a count), name,
    and release data. it adds 25 items at a time because that is how many is returned by episodes_search'''

def createPieChart(cur):
    '''Once the above 3 functions are run and an appropriate amount of data is gathered, run this function to
    create a pie chart that shows the percentage of 'Special Episodes.' These are episodes that are not numbered
    with Joe Rogan's normal numbering sequence. It uses the Spotify Episodes table in JRP.db to get the data
    and counts how many episodes start with a # (which is his 'normal' numbering sequence) how many don't'''

def createBarGraph(cur, file):
    '''this function, which should be run at the same time as createPieChart, creates a bar graph showing the number
    of episodes released in each month of 2020. It uses Spotify Episodes and splits the release date into its respective
    parts, which are then used to create counts for each month. It then displays these findings in a bar graph AND
    writes them to a text file, which can be named with the 'file' variable.'''

def main():
    '''this function is used to call the above functions. it is recommended to call the first three at once
    multiple times, and then the last two only once.'''
```



Date	Issue Description	Location of Resource	Result
12/2/20	Needed to start reading a csv file after a certain row(imitating the 25 upload maximum of an API)	https://stackoverflow.com/question/s/26464567/csv-read-specific-row	Resource allowed me to use a csv reader and next() function to get a row containing a specific row number
12/1/20	Needed to change a specific cell in SQL database to put a shared id value for a whole column	https://stackoverflow.com/questions/3024546/change-one-cells-data-in-mysql	Using a loop and (UPDATE my_table SET my_column='new value' WHERE something='some value')I was able to change a whole column values to s shared guestid
12/5/20	Y axis of 'views' for a bar graph was being put in exponential form and it was not intuitive to look at	https://stackoverflow.com/questions/14711655/how-to-prevent-numbers-being-changed-to-exponential-form-in-python-matplotlib-fi	Showed how to change the ticklabels to useOffset=False which shows the views in by pure millions instead
12/5/20	X axis names of a bar chart were crossing over each other making it hard to read	https://stackoverflow.com/question/s/43152502/how-can-i-rotate-xticklabels-in-matplotlib-so-that-the-spacing-between-each-xtic	Showed how to rotate the x-axis labels by 45 and make the font smaller to make it easier to read with <code>plt.setp(ax.get_xticklabels(), ha="right", rotation=45)</code>



Date	Issue Description	Location of Resource	Result
12/5/20	X-axis labels were long, cutoff and unreadable even when shifted 45 degrees	https://stackoverflow.com/question/s/59466109/how-to-get-x-axis-labels-in-multiple-line-in-matplotlib	Resource allowed me to use a csv reader and next() function to get a row containing a specific row number
12/1/20	Needed to change a specific cell in SQL database to put a shared id value for a whole column	https://stackoverflow.com/questions/3024546/change-one-cells-data-in-mysql	Using a loop and (UPDATE my_table SET my_column='new value' WHERE something='some value') I was able to change a whole column values to s shared guestid
12/5/20	Y axis of 'views' for a bar graph was being put in exponential form and it was not intuitive to look at	https://stackoverflow.com/questions/14711655/how-to-prevent-numbers-being-changed-to-exponential-form-in-python-matplotlib-fi	Shown how to change the ticklabels to useOffset=False which shows the views in by pure millions instead
12/5/20	X axis names of a bar chart were crossing over each other making it hard to read	https://stackoverflow.com/question/s/43152502/how-can-i-rotate-xticklabels-in-matplotlib-so-that-the-spacing-between-each-xtic	Shown how to rotate the x-axis labels by 45 and make the font smaller to make it easier to read with plt.setp(ax.get_xticklabels(), ha="right", rotation=45)