NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
UNIVERSITY OF PELOPONNESE

MSc in Data Science
2021-2022

Big Data Management
Assignment 2

REPORT

Vitsas Alexandros-Konstantinos

Spiliakos Georgios

Contents

1 Data Preprocessing

For the data pre-processing we used Python Pandas library as it has build in functions ideal for processing data and MongoDB Query Language. The process we followed is shown below:

- Load .csv: We loaded the .csv files as Pandas dataframes.
- Save .json: We saved the dataframe as .json.
- Process ’trending_date’: We used MongoDB query language in order to process the the ’trending_date’ column and make it in the correct date format. Specifically we concated number ’20’ in front of the ’year’ part of the date and then used ’dateFromString’ to reformat it.
- Process ’publish_time’: We used MongoDB query language in order to process the the ’trending_date’ using ’dateFromString’ to reformat it.
- Process ’tags’: We used MongDB query language in order to create an Array of objects instead of a Strings, using ’split’ function with "|" separator.
- The data were loaded into two different tables namely: videosGB, videosUS.

2 Tools

Tools that were used are:

- Docker: In order to easily create and start/stop our Mongo database
- pymongo: In order to connect and use our MongoDB from our Python script using MongoClient.
- MongoDB query language: In order to process the data and execute the queries.
- Pandas: Used for procedural processes like save the data and the query results.
- matplotlib: In order to create the plots required for the results.

We also used MongoDB Compass which had the useful ’Pipeline’ feature in which we initially executed the queries. Pymongo was chosen afterwards to easily share and execute the queries.

3 Queries

3.1 Preprocessing queries

The preprocessing queries used ’update_many’ to the tables.

3.1.1 Trending Date pre-processing

```
{
    '$set': {
        'trending_date': {
            '$concat': [
                '20', '$trending_date'
            ]
        }
    }
}, {
    '$set': {
        'trending_date': {
            '$dateFromString': {
                'dateString': '$trending_date',
                'format': '%Y.%d.%m'
            }
        }
    }
}
```

3.1.2 Publish time pre-processing

```
{
    '$set': {
        'publish_time': {
            '$dateFromString': {
                'dateString': '$publish_time'
            }
        }
    }
}
```

3.1.3 Tags pre-processing

```
{
  "$addFields":{
     "tags":{
        "$replaceAll":{
           "input":"$tags",
           "find":"\"",
           "replacement":""
        }
     }
  }
},
{
  "$addFields":{
     "tags":{
        "$split":[
           "$tags",
           "|"
        ]
     }
  }
}
```

3.2 Query 1

3.2.1 Query 1

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $match: {
  channel_title: 'Saturday␣Night␣Live'
 }
}, {
 $project: {
  title: true,
  views: true,
  likes: true,
  dislikes: true
 }
}, {
 $sort: {
  views: −1
 }
}]
```

3.2.2 Query 1 results

```
{"title":"Royal Wedding - SNL","views":8607264,"likes":66559,"dislikes":14179},
{"title":"A Kanye Place - SNL","views":5547578,"likes":110621,"dislikes":5101},
{"title":"Natalie\u2019s Rap 2 - SNL","views":5156609,"likes":79865,"dislikes":3195},
{"title":"George W. Bush Returns Cold Open - SNL","views":5147621,"likes":55076,"dislikes":6752},
{"title":"Welcome to Hell - SNL","views":4649310,"likes":60641,"dislikes":8997},
{"title":"Meet the Parents Cold Open - SNL","views":4548677,"likes":40718,"dislikes":4666},
{"title":"Morning Joe Michael Wolff Cold Open - SNL","views":4162540,"likes":31716,"dislikes":3658},
{"title":"Presidential Address Cold Open - SNL","views":3763816,"likes":31791,"dislikes":3603},
{"title":"What Even Matters Anymore - SNL","views":3362428,"likes":35137,"dislikes":10550},
{"title":"Visit with Santa Cold Open - SNL","views":2935898,"likes":28777,"dislikes":4642},
{"title":"White House Tree Trimming Cold Open - SNL","views":2701442,"likes":27542,"dislikes":5428},
{"title":"Eminem: Walk on Water, Stan, Love the Way You Lie (ft. Skylar Grey) (Live) - SNL","views":2200854,"likes":36494,"dislikes":3044},
{"title":"Weekend Update: Stefon on St. Patrick's Day - SNL","views":2189861,"likes":23694,"dislikes":761},
{"title":"Weekend Update on Donald Trump's Asia Trip - SNL","views":1907912,"likes":18641,"dislikes":1741},
{"title":"Queer Eye's Tan France Takes Pete Davidson Shopping - SNL","views":1856381,"likes":32619,"dislikes":1483},
{"title":"Taylor Swift: \u2026Ready for It? (Live) - SNL","views":1776127,"likes":35741,"dislikes":3214},
{"title":"Family Feud: Oscars Edition - SNL","views":1352472,"likes":16328,"dislikes":857},
{"title":"Handmaids in the City - SNL","views":1250529,"likes":9870,"dislikes":2437},
{"title":"Cut for Time: New Year\u2019s Kiss - SNL","views":830143,"likes":10478,"dislikes":871},
{"title":"SNL Host Donald Glover Is Not Here for Beck Bennett's Tribute","views":411490,"likes":9515,"dislikes":259}
```

3.2.3 Query 1 comments

The channel Saturday Night Live seems to be a very popular channel covering famous events and persons as it can be seen from their title and amount of views. Despite this fact we are able to see a low amount of user interaction as the amount of likes and dislikes is nonequivalent to the amount of views, remaining in low thousands while the views are several thousands or millions. Also we are able to see a positive acceptance from viewers as likes are more than dislikes.

3.3 Query 2

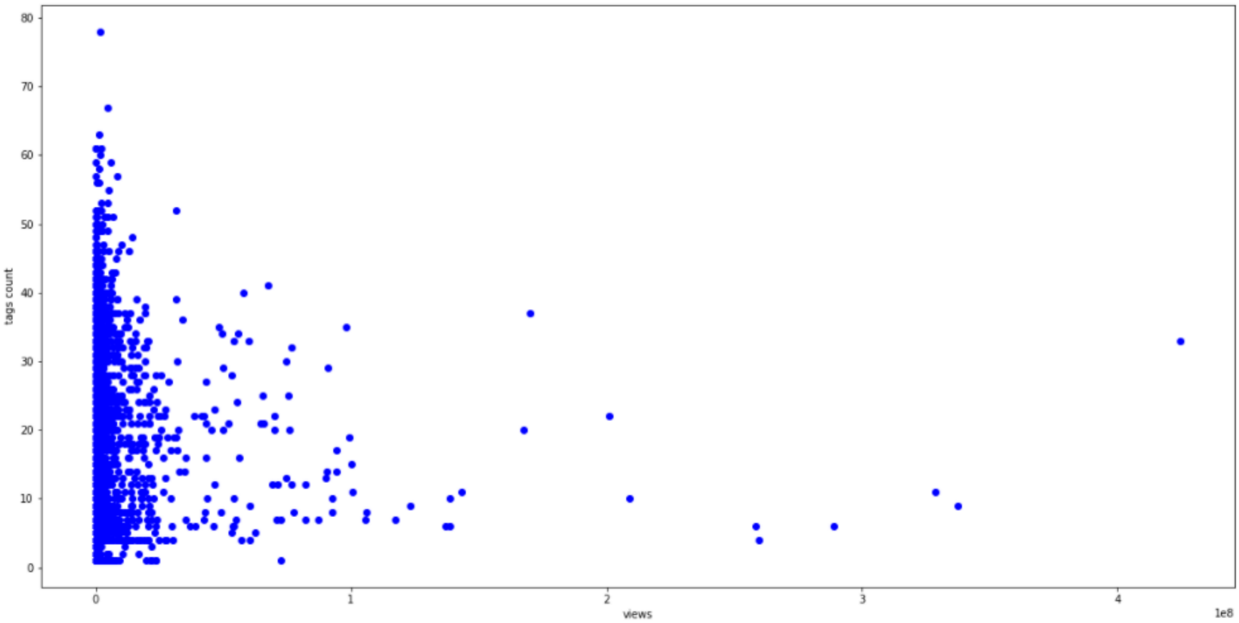
3.3.1 Query 2

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $project: {
  video_id: true,
  tags: {
   $size: '$tags'
  },
  views: true
 }
}, {
 $sort: {
  views: −1
 }
}]
```

3.3.2 Query 2 results

```
{"video_id":"_I_D_8Z4sJE","views":424538912,"tags":33},
{"video_id":"9jI-z9QN6g8","views":337621571,"tags":9},
{"video_id":"kLpH1nSLJSs","views":328860380,"tags":11},
{"video_id":"wfWkmURBNv8","views":288811992,"tags":6},
{"video_id":"VYOjWnS4cMY","views":259721696,"tags":4},
{"video_id":"xpVfcZ0ZcFM","views":258164991,"tags":6},
{"video_id":"ffxKSjUwKdU","views":208876887,"tags":10},
{"video_id":"zEf423kYfqk","views":200862743,"tags":22},
{"video_id":"FlsCjmMhFmw","views":169884583,"tags":37},
{"video_id":"sGIm0-dQd8M","views":167456025,"tags":20},
{"video_id":"TyHvyGVs42U","views":143408235,"tags":11},
{"video_id":"2Vv-BfVoq4g","views":138578860,"tags":10},
{"video_id":"M4ZoCHID9GI","views":138535053,"tags":6},
{"video_id":"Ck4xHocysLw","views":137081637,"tags":6},
{"video_id":"7C2z4GqqS5E","views":123010920,"tags":9},
{"video_id":"tCXGJQYZ9JA","views":117270304,"tags":7},
{"video_id":"U9BwWKXjVaI","views":106147032,"tags":8},
{"video_id":"au2n7VVGv_c","views":105629911,"tags":7},
{"video_id":"6ZfuNTqbHE8","views":100672931,"tags":11},
{"video_id":"fGqdIPer-ms","views":100159686,"tags":15}
```

3.3.3 Query 2 plot



3.3.4 Query 2 comments

We are able to see that the videos that have extreme viewership are using around 10-40 tags with the big amount using <10 tags. In general the videos that have a good performance seem to be using <20 tags. Interestingly enough we can see that there is an amount of videos that with 0 tags are able to perform decently, while videos with >40 are performing bad.

3.4 Query 3

3.4.1 Query 3

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $unwind: {
  path: '$tags'
 }
}, {
 $group: {
  _id: '$tags',
  count_of_video: {
   $sum: 1
  }
 }
}, {
 $sort: {
  count_of_video: -1
 }
}]
```

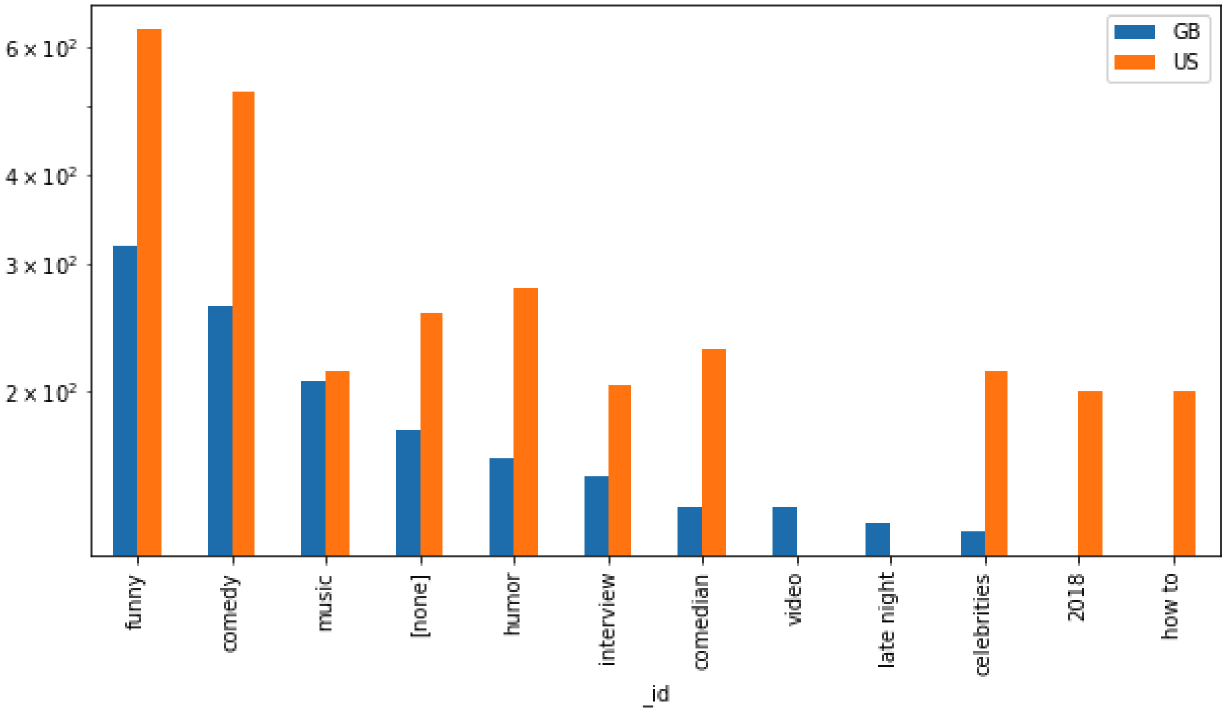
3.4.2 Query 3 results for GB

```
{"_id":"funny","count_of_video":318},
{"_id":"comedy","count_of_video":263},
{"_id":"music","count_of_video":206},
{"_id":"[none]","count_of_video":176},
{"_id":"humor","count_of_video":161},
{"_id":"interview","count_of_video":152},
{"_id":"video","count_of_video":138},
{"_id":"comedian","count_of_video":138},
{"_id":"late night","count_of_video":131},
{"_id":"celebrities","count_of_video":128},
{"_id":"funny video","count_of_video":126},
{"_id":"jokes","count_of_video":126},
{"_id":"live","count_of_video":117},
{"_id":"2018","count_of_video":113},
{"_id":"hollywood","count_of_video":106},
{"_id":"celebrity","count_of_video":103},
{"_id":"clip","count_of_video":101},
{"_id":"show","count_of_video":99},
{"_id":"comedic","count_of_video":96},
{"_id":"Pop","count_of_video":92}
```

3.4.3 Query 3 results for US

```
{"_id":"funny","count_of_video":636},
{"_id":"comedy","count_of_video":521},
{"_id":"humor","count_of_video":278},
{"_id":"[none]","count_of_video":258},
{"_id":"comedian","count_of_video":229},
{"_id":"celebrities","count_of_video":213},
{"_id":"music","count_of_video":213},
{"_id":"interview","count_of_video":204},
{"_id":"2018","count_of_video":200},
{"_id":"how to","count_of_video":200},
{"_id":"celebrity","count_of_video":197},
{"_id":"funny video","count_of_video":196},
{"_id":"video","count_of_video":185},
{"_id":"jokes","count_of_video":184},
{"_id":"news","count_of_video":178},
{"_id":"food","count_of_video":174},
{"_id":"science","count_of_video":173},
{"_id":"late night","count_of_video":167},
{"_id":"NBC","count_of_video":164},
{"_id":"live","count_of_video":162}
```

3.4.4 Query 3 Plot



3.4.5 Query 3 comments

We are able to see that there are common tags used by both GB and US countries such as ’funny’, ’comedy’ and ’music’. In most cases the tags used in US are used in double the amount of videos in contrast to the GB. Also we are able to see that the most used tags revolve around humorous videos. Lastly a big amount of videos did not use any tags which happens in both countries.

3.5 Query 4

3.5.1 Query 4 for comments_disabled:true

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $match: {
  comments_disabled: true
 }
}, {
 $group: {
  _id: null,
  avg_views: {
   $avg: '$views'
  },
  avg_likes: {
   $avg: '$likes'
  },
  avg_dislikes: {
   $avg: '$dislikes'
  }
 }
}]
```

3.5.2 Query 4 for comments_disabled:false

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $match: {
  comments_disabled: false
 }
}, {
 $group: {
  _id: null,
  avg_views: {
   $avg: '$views'
  },
  avg_likes: {
   $avg: '$likes'
  },
  avg_dislikes: {
   $avg: '$dislikes'
```

```
    }
  }
}]
```

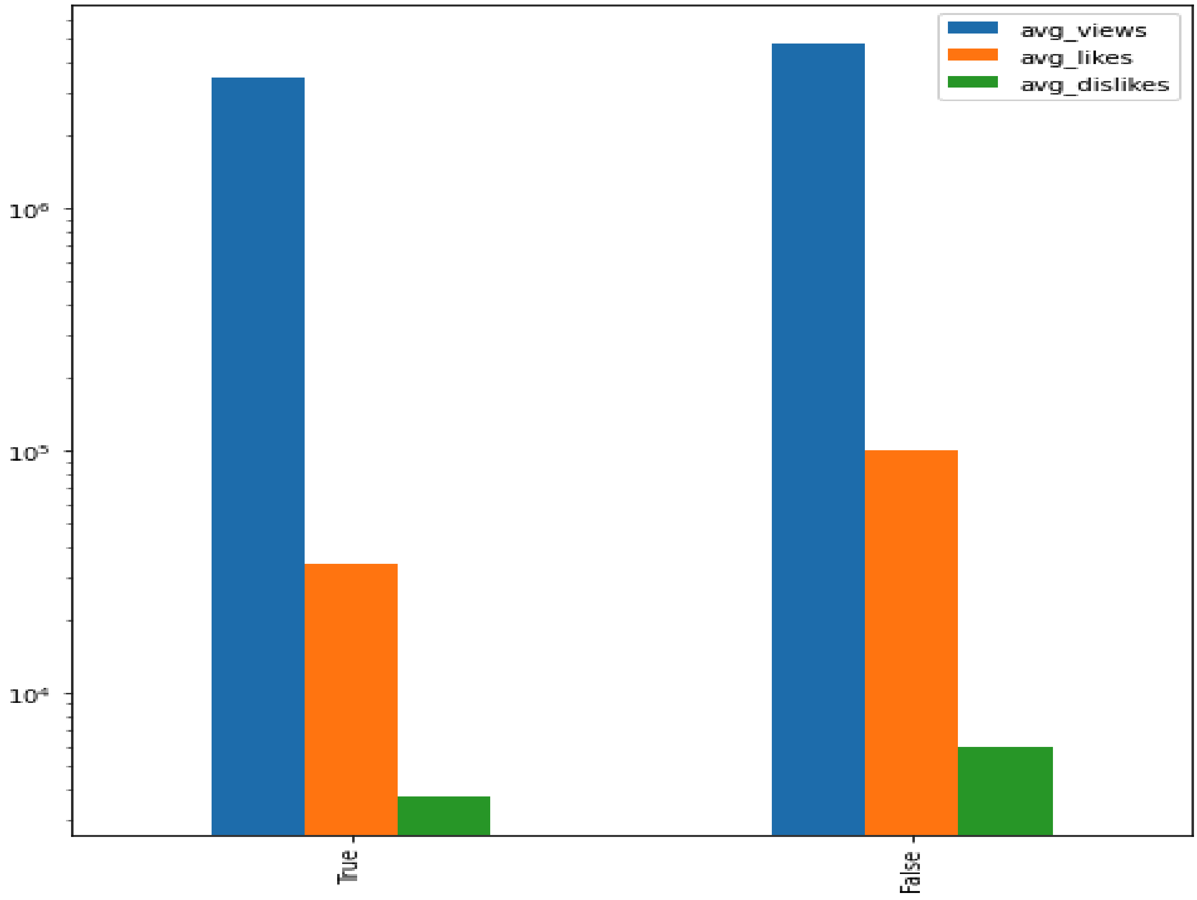
3.5.3 Query 4 results for comments_disabled:true

```
 {"avg_views":3488488.3137254901, "avg_likes":33803.6470588235, "avg_dislikes"
    :3716.3137254902}
```

3.5.4 Query 4 results for comments_disabled:false

```
{"avg_views":4835556.6435889471, "avg_likes":100914.2719652282, "avg_dislikes"
   :5944.5768394908}
```

3.5.5 Query 4 Plot



3.5.6 Query 4 comments

In regards to viewership we are able to see that there is no big difference in both cases. In regards to likes and dislikes we are able to see that when a video offers comments has more interaction via the buttons. Lastly we assumed that videos with comments disabled would have bigger usage in likes/dislikes buttons as the viewers would have reduced ways to interact with the video, however the reality seems to be the opposite.

3.6 Query 5

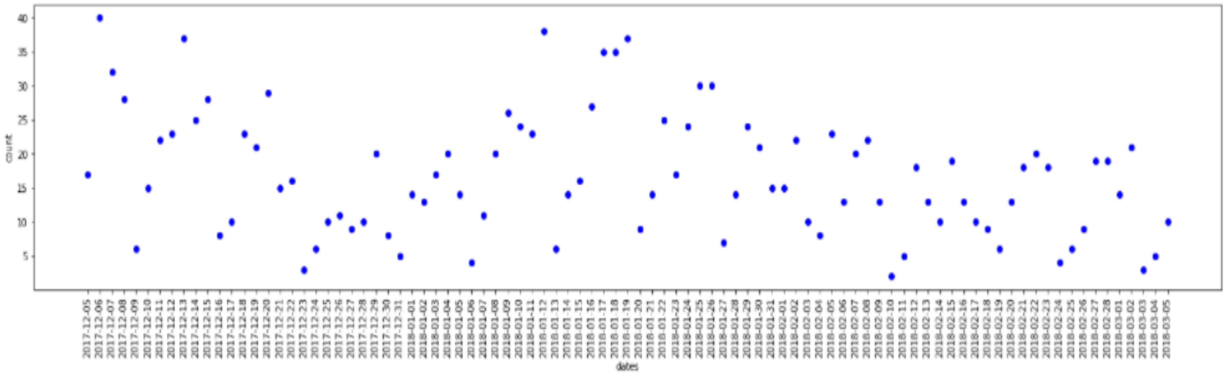
3.6.1 Query 5

```
[{
 $sort: {
  views: 1
 }
}, {
 $group: {
  _id: '$video_id',
  last: {
   $last: '$$ROOT'
  }
 }
}, {
 $replaceRoot: {
  newRoot: '$last'
 }
}, {
 $match: {
  publish_time: {
   $gte: ISODate('2017-12-05T00:00:00.000Z'),
   $lte: ISODate('2018-03-05T23:59:59.999Z')
  }
 }
}, {
 $group: {
  _id: {
   $dateToString: {
    format: '%Y-%m-%d',
    date: '$publish_time'
   }
  },
  count: {
   $sum: 1
  },
  date: {
   $first: '$publish_time'
  }
 }
}, {
 $sort: {
  date: 1
 }
}, {
 $addFields: {
  date: '$_id'
 }
}, {
 $project: {
  _id: false
 }
}]
```

3.6.2 Query 5

```
{"count":17, "date":"2017-12-05"},
{"count":40, "date":"2017-12-06"},
{"count":32, "date":"2017-12-07"},
{"count":28, "date":"2017-12-08"},
{"count":6, "date":"2017-12-09"},
{"count":15, "date":"2017-12-10"},
{"count":22, "date":"2017-12-11"},
{"count":23, "date":"2017-12-12"},
{"count":37, "date":"2017-12-13"},
{"count":25, "date":"2017-12-14"},
{"count":28, "date":"2017-12-15"},
{"count":8, "date":"2017-12-16"},
{"count":10, "date":"2017-12-17"},
{"count":23, "date":"2017-12-18"},
{"count":21, "date":"2017-12-19"},
{"count":29, "date":"2017-12-20"},
{"count":15, "date":"2017-12-21"},
{"count":16, "date":"2017-12-22"},
{"count":3, "date":"2017-12-23"},
{"count":6, "date":"2017-12-24"}
```

3.6.3 Query 5 Plot



3.6.4 Query 5 comments

We are able to see that there is a higher video posting in early December and early January. There is a lower video posting during Christmas holidays and a higher count of video postings on January. Also Thursdays and Fridays seems common in having higher video postings probably aiming for weekend views while viewers have more free time. Lastly weekends are have a lower amount of video postings across all 3 months.

4 User Manual

4.0.1 Prerequisites

- Python 3.9
- Jupyter Notebooks
- Docker

4.0.2 Steps

- docker compose up —d
- Open a virtual environment with Python 3.9
- Use pip3 install —r requirements.txt to install the required libraries.
- Open Jupyter Server.
- Execute the Jupyter Notebook sequentially.
- The results will be shown in the ’results’ directory.