**A Project Report**

**On**

***NETFLIX DATA ANALYSIS***

Submitted by:

BISHAL SAIKIA

Under the guidance of

**Mr. Prabhat Das**

(PGT-IP)



**ARMY PUBLIC SCHOOL JORHAT**

**MES GATE ROWRIAH JORHAT**

**Assam-785004**

**CERTIFICATE BY PRINCIPAL**

This is to certify that this project report entitled “Netflix Data Analysis” submitted by Bishal Saikia to Army Public School Jorhat has been examined and evaluated.

The report has been prepared as per the regulations of CBSE and qualifies to be accepted.

Date:

Place:

Mrs. Ferdausi Sultana Hazarika

(Principal)

Army Public School Jorhat

**CERTIFICATE BY EXAMINERS**

This is to certify that this project report entitled “**NETFLIX DATA ANALYSIS**” is the bonafide work of **BISHAL SAIKIA** who carried out the project work under my supervision and guidance.

To the best of my knowledge, the matter embodied in the report has not been submitted to any other institute for the award of any other degree.

Date:

Place:

Mr. Prabhat Das

(External Examiner) (InternalExaminer)

**ACKNOWLEDGEMENT**

I take this opportunity to extend my heart full gratitude to Army Public School Jorhat for providing me the opportunity.

I am highly grateful to my guide Mr. Prabhat Das, PGT-IP, Army Public School Jorhat for giving us the opportunity to work under him and providing us an ample guidance and support through the project.

Lastly, I would also like to thank the authors whose publications guided us regarding our project.

**DECLARATION**

I admit that this report is of my own work and all the sources of the information used in this report have fully acknowledged.

I hereby declare that the dissertation work entitled “**NETFLIX DATA ANALYSIS**” submitted to the Army Public School Jorhat, is prepared by me and was not submitted to any other institution for award of any other degree.

Date:

Place:

Signature

**Abstract**

Netflix is one of the largest providers of online streaming services. It collects a huge amount of data because it has a very large subscriber base. In this project, we’re going to introduce you to a data science project on Netflix data analysis. We can analyse a lot of data and models from Netflix because this platform has consistently focused on changing business needs by shifting its business model from on-demand DVD movie rental and now focusing a lot about the production of their original shows. Some of the important tasks that we analysed from Netflix data are: Sentiment analysis of content available on Netflix, contents produced in a particular year, Runtime of different movies available on Netflix, their release Years etc.

The dataset we used for the Netflix data analytics task consists of movies streamed on Netflix of all time. The dataset is provided by Kaggle which allows users to find and publish data sets, explore and build models in a web-based data-science environment.

**CONTENTS**

                                                                                  Page No.

1. Tools and libraries used                                                    1
2. Introduction                                                                     3
3. Project Overview                                                              4
4. Source code                                                                      9
5. Commands used in MySQL                                             16
6. Conclusion and future work                                             17
7. References                                                                        18

**List of Figures**

Page no

* Fig 1. User information table 4
* Fig 2.1. Login options 4
* Fig 3.1. Login for registered user 4
* Fig 3.2. Login for new user 5
* Fig 4. Output options 5
* Fig 5. Output for option 1 5
* Fig 6. Output for option 2 6
* Fig 7. Output for option 3 6
* Fig 8. Output for option 4 6
* Fig 9. Output for option 5 6
* Fig 10. Output for option 6 7
* Fig 11. Output for option 7 8

**Tools and Libraries Used**

MySQL

MySQL is an open-source relational database management system (RDBMS). Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language, programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

MySQL is free and open-source software under the terms of the GNU General Public License, and is also available under a variety of proprietary licenses. MySQL was owned and sponsored by the Swedish company MySQL AB, which was bought by Sun Microsystems (now Oracle Corporation). In 2010, when Oracle acquired Sun, Widenius forked the open-source MySQL project to create MariaDB [1].

PyCharm

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

PyCharm is cross-platform, with Windows, macOS and Linux versions.

The Community Edition is released under the Apache License, and there is also Professional Edition with extra features – released under a proprietary license [2].

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010 [3].

Text Blob

TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, Netflix Data Analysis, and more [4].

Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

Matplotlib was originally written by John D. Hunter. Since then it has an active development community and is distributed under a BSD-style license [5].

**Introduction**

Netflix started in 1997 as a rent-by-mail DVD service that used a pay-per-rental model. Users would browse and order the films they wanted on their website, put in an order, and Netflix would post them to your door. After renters had finished with the DVDs, they would simply post them back. On July 10, 2020, Netflix became the largest entertainment company by market capitalization. The company has grown from a DVD mail order format to being the king of original content. With the impact of Covid-19 on the operation of movie theatres, and more people having to remain indoors, Netflix may soon become ‘the’ viable alternative

With over 203.67 million subscribers worldwide, Netflix has reigned as the #1 Over-the-Top (OTT) streaming content platform since its launch in 2007. The following program is an attempt to develop a program that can analyse the data of movies available on Netflix from a given dataset and assign its rating. It includes a simple registration system where the user can register him/her into the program with the help of a one-time password (OTP) sent to the email address used for the registration process. The program then gives us a number of options to get different information about the movie entered present in the given dataset. A graph is also generated using the Matplotlib library that gives us a visual representation of the polarity of the given data.

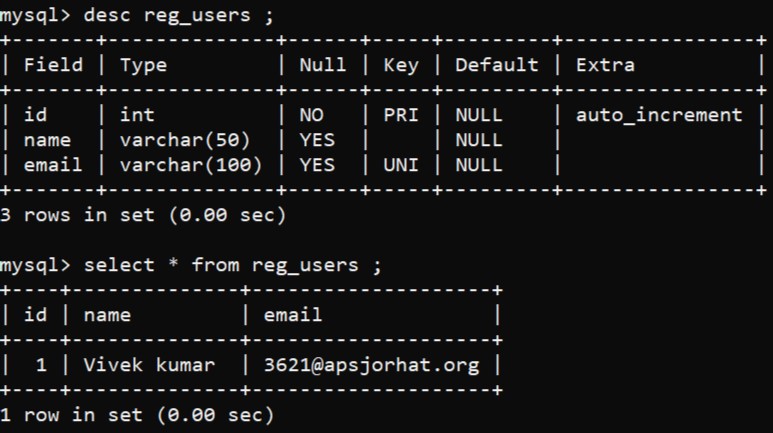
**Project Overview**

Figure 1: All the information regarding the users is stored in this table

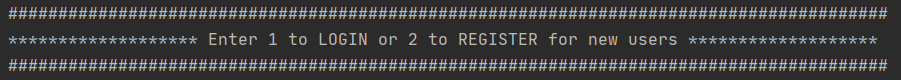


Figure 2 : When the home page is opened two login options are presented to the user

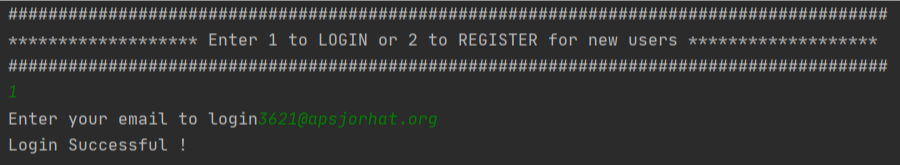


Figure 3.1: When the user chooses option 1 the program asks for his/her email address, on entering the email id, if the email id exists then the login procedure is completed.

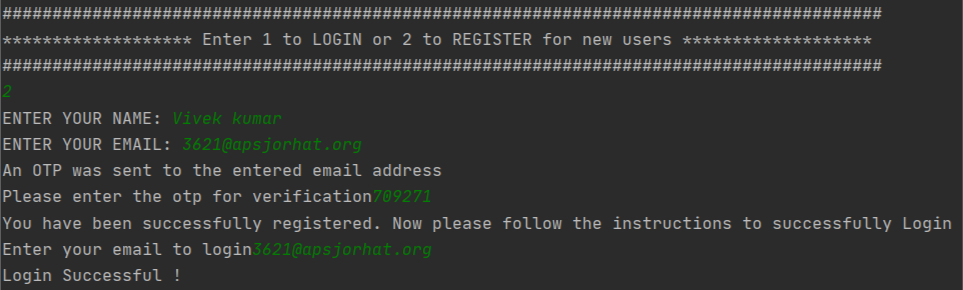


Figure 3.2 Login procedure for new users

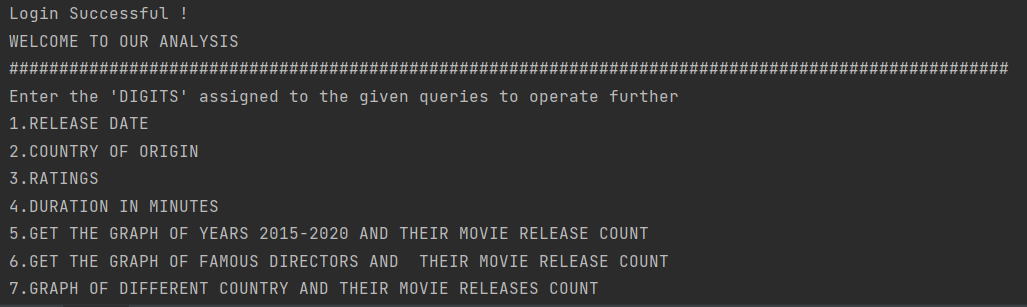


Figure 4: After the login procedure is completed the user is presented with the following list of options

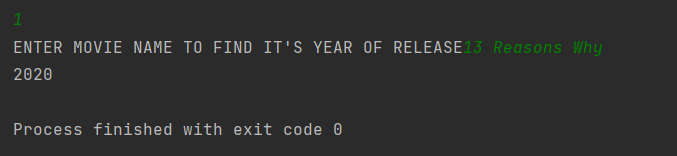


Figure 5: Output for option 1 (2020)

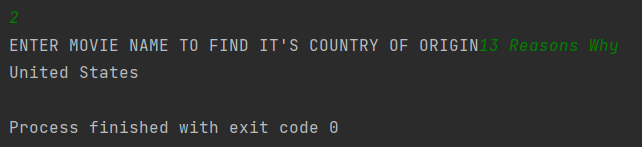


Figure 6: Output for option 2 (United States)

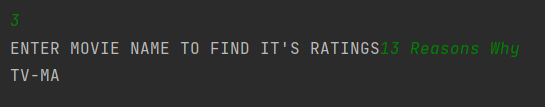


Figure 7: Output for option 3 (TV-MA is the rating that shows the program is intended for adults.)

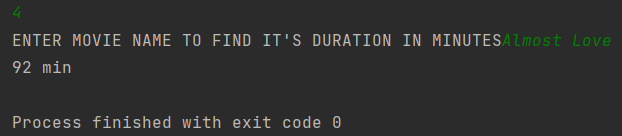
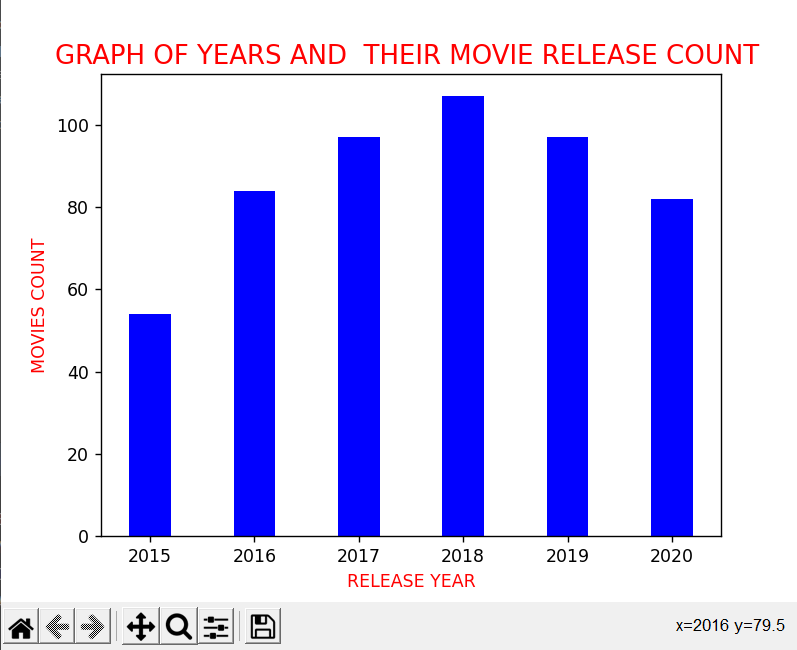


Figure 8: Output for option 4 (92minutes)



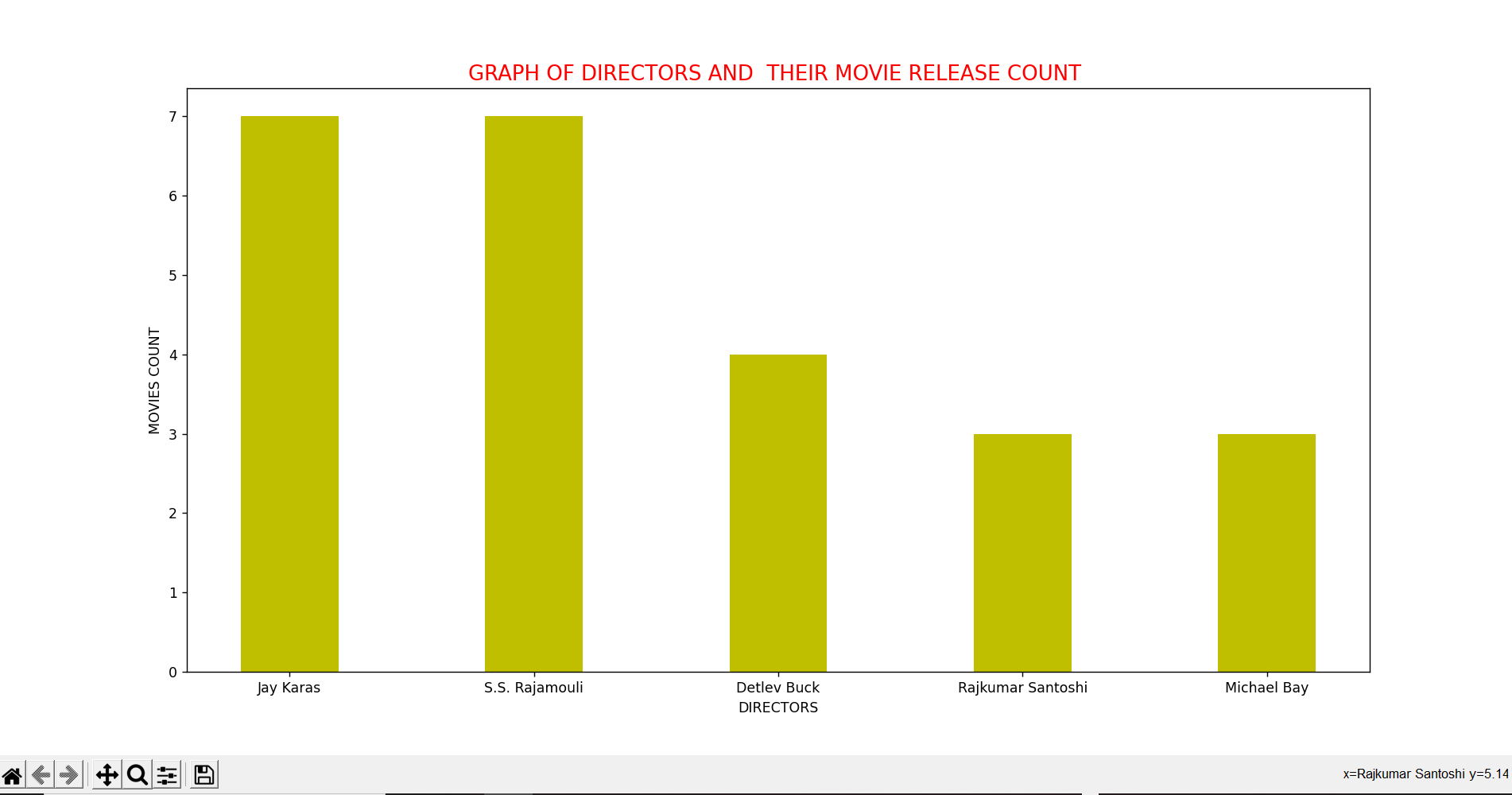
Figure 9: Output for option 5 (Graph)

Figure 10: Output for option 6 (Graph

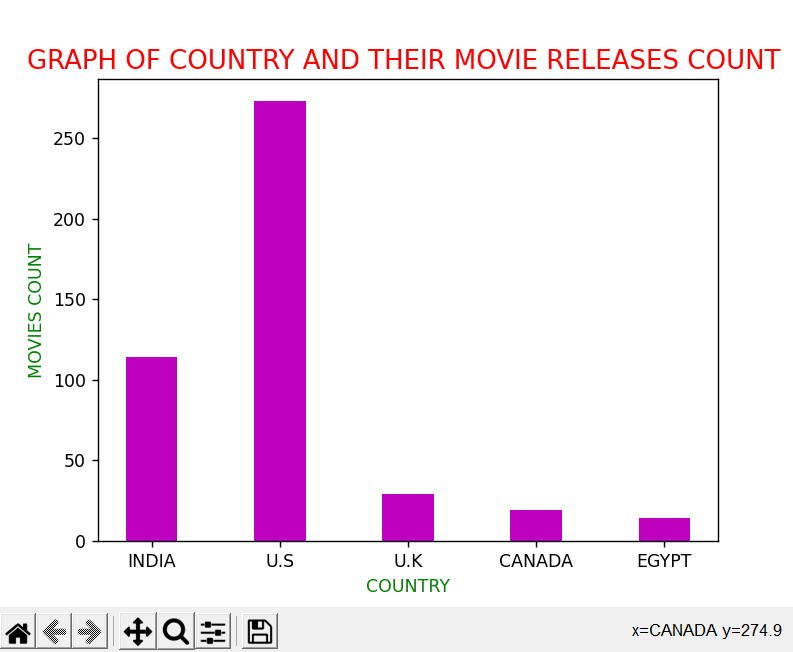


Figure 11: Output for option 7 (Graph)

**Source Code**

**MAIN**

import mysql.connector as db

import smtplib

import random

mydb=db.connect(host="localhost",passwd="root",user="root",database="netflix\_data\_analysis")

mycursor=mydb.cursor()

def user\_auth():

count=None

email = input("Enter your email to login")

fetch\_query = "select \* from reg\_users;"

mycursor.execute(fetch\_query)

for i in mycursor:

if i[2] == email:

print("Login Successful !")

import dashboard

count = 1

break

if count == 0:

print("Login Error !")

print("#"\*88)

print("\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Enter 1 to LOGIN or 2 to REGISTER for new users \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*")

print("#"\*88)

inp=int(input())

count=0

if inp==1:

user\_auth()

elif inp==2:

name = input("ENTER YOUR NAME: ")

email = input("ENTER YOUR EMAIL: ")

otp = str(random.randint(100000,1000000))

SUBJECT = "OTP for Login"

TEXT = "HEY "+name+'!'"\r\n""\r\n"'Your OTP for login is: ' + otp+ "\r\n""\r\n"'Please enter the otp for further verification'"\r\n""\r\n"'Thank you !'

s = smtplib.SMTP('smtp.gmail.com', 587)

s.starttls()

s.login('group2@apsjorhat.org', 'apsj#12345678')

message = 'Subject: {}\n\n{}'.format(SUBJECT, TEXT)

s.sendmail('group2@apsjorhat.org', email, message)

s.quit()

print("An OTP was sent to the entered email address")

answer=int(input("Please enter the otp for verification"))

if answer == int(otp) :

query = "insert into reg\_users (name,email) values (" + "'" + name + "'"+"," + "'" + email + "'" + ");"

mycursor.execute(query)

mydb.commit()

print("You have been successfully registered. Now please follow the instructions to successfully Login ")

user\_auth()

else:

print('Login failed. Please Try Again')

**FUNCTIONS**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import matplotlib.pyplot as plt

df=pd.read\_csv("C:/Users/borgo/PycharmProjects/APSJ21\_G2\_NETFLIX\_DATA\_ANALYSIS/dataset.csv",engine="python")

years=df['country'].value\_counts()

def movie\_name():

a=str(input("ENTER MOVIE NAME TO FIND IT'S YEAR OF RELEASE"))

details=df[df['title']==a]

for i in details['release date']:

print(i)

def movie\_country():

b=str(input("ENTER MOVIE NAME TO FIND IT'S COUNTRY OF ORIGIN"))

details=df[df['title']==b]

for i in details['country']:

print(i)

def movie\_ratings():

c=str(input("ENTER MOVIE NAME TO FIND IT'S RATINGS"))

details=df[df['title']==c]

for i in details['ratings']:

print(i)

def movie\_duration():

d=str(input("ENTER MOVIE NAME TO FIND IT'S DURATION IN MINUTES"))

details=df[df['title']==d]

for i in details['duration(min)']:

print(i,'min')

def yvm\_graph():

ndy=['2015','2016','2017','2018','2019','2020']

ndc=[54,84,97,107,97,82]

plt.bar(ndy, ndc, width=0.4,color='b')

plt.ylabel('MOVIES COUNT', size=10, color='r')

plt.xlabel('RELEASE YEAR', size=10, color='r')

plt.title('GRAPH OF YEARS AND THEIR MOVIE RELEASE COUNT ', size=15, color='r')

plt.show()

def dvm\_graph():

ndd=['Jay Karas','S.S. Rajamouli','Detlev Buck','Rajkumar Santoshi','Michael Bay']

ndm=[7,7,4,3,3]

plt.figure(figsize=(12, 7))

plt.bar(ndd, ndm, width=0.4,color='y')

plt.ylabel('MOVIES COUNT', size=10, color='k')

plt.xlabel('DIRECTORS', size=10, color='k')

plt.title('GRAPH OF DIRECTORS AND THEIR MOVIE RELEASE COUNT ', size=15, color='r')

plt.show()

def cvm\_graph():

cn=['INDIA','U.S','U.K','CANADA','EGYPT']

cnc=[114,273,29,19,14]

plt.bar(cn, cnc, width=0.4,color='m')

plt.ylabel('MOVIES COUNT', size=10, color='g')

plt.xlabel('COUNTRY', size=10, color='g')

plt.title('GRAPH OF COUNTRY AND THEIR MOVIE RELEASES COUNT ', size=15, color='r')

plt.show()

**DASHBOARD**

import textblob as tb

import matplotlib.pyplot as plt

import numpy as np

import functions

def dashboard():

print("WELCOME TO OUR ANALYSIS")

print('#'\*100)

print("Enter the 'DIGITS' assigned to the given queries to operate further\n"

"1.RELEASE YEAR \n"

"2.COUNTRY OF ORIGIN \n"

"3.RATINGS \n"

"4.DURATION IN MINUTES \n"

"5.GET THE GRAPH OF YEARS 2015-2020 AND THEIR MOVIE RELEASE COUNT \n"

"6.GET THE GRAPH OF FAMOUS DIRECTORS AND THEIR MOVIE RELEASE COUNT \n"

"7.GRAPH OF DIFFERENT COUNTRY AND THEIR MOVIE RELEASES COUNT \n")

"\n"

inp=int(input())

if inp==1:

functions.movie\_name()

if inp==2:

functions.movie\_country()

if inp==3:

functions.movie\_ratings()

if inp==4:

functions.movie\_duration()

if inp==5:

functions.yvm\_graph()

if inp==6:

functions.dvm\_graph()

if inp==7:

functions.cvm\_graph()

dashboard()

**COMMANDS USED IN MySQL**

**Creating database**

Create database netflix\_data\_analysis;

**Using database**

Use netflix\_data\_analysis**;**

**Creating table and inserting values**

Create table reg\_users(id int(10) primary key auto\_increment ,name varchar(100),email varchar(100) unique);

**Desc table**

Desc reg\_users**;**

**Inserting values**

insert into reg\_users values((1,"Bishal Saikia ", "4010@apsjorhat.org"),(2,"Frachen Borgohain","3901@apsjorhat.org"),(3,”Himanshu Gogoi","5434@apsjorhat")) ;

**To fetch all values**

Select \* from reg\_users;

**Conclusion and Future Work**

In this project we have created a program that can be used to analyse movies available in Netflix for details such as their release year, duration and rating.

Most of the world’s data is unstructured and unorganised and a program such as ours can help sort the data in an efficient manner and generate useful information regarding movies available in Netflix.

Netflix Data Analysis is important because it helps analyse the growth in production of movies and also the expanding business of Netflix. By automatically sorting their ratings we can classify the movies in different categories appropriate for different age groups.

Netflix has grown into one of the most used video streaming services in the world. The company was one of the first to see the potential of video streaming technology and began to transition to a subscription video-on-demand model in 2007. Since this transition, Netflix’s revenue has grown from 1.36 billion to around 25 billion in just 12 years. The number of Netflix subscribers has followed a similar trend, increasing from less than 22 million in 2011 to nearly 214 million in 2021. What we have developed here is an elementary Netflix Data Analysis program.

There is a lot of scope for upgrading our program, we can use advanced Netflix Data Analysis models that are more accurate and can provide more information about movies as well as TV shows. We can also add the functionality to link our program with social media sites for Netflix Data Analysis and compare its popularity with other popular OTT platforms. There is also the scope to develop a web-based application for Netflix Data Analysis.

**References**

[1] <https://en.wikipedia.org/wiki/MySQL>

[2] <https://en.wikipedia.org/wiki/PyCharm>

[3] <https://en.wikipedia.org/wiki/Pandas_(software)>

[4 ]<https://textblob.readthedocs.io/en/dev/>

[5] <https://en.wikipedia.org/wiki/Matplotlib>