

**SHRI VAISHNAV VIDHYAPEETH VISHWAVIDYALAYA**

**INDORE**

**DATA SCIENCE**

**USING PYTHON**

**“NETFLIX VISUALIZATION AND RECOMMENDATION SYSTEM”**

**TRAINING PROJECT REPORT**

**SUBMITTED BY:**

ALINA KHAN

**1710DMTCSE01286**

SHEFALI PARMAR

**1710DMTCECF01333**

PRAPTI PAGNIS

**1710DMTCSE01303**

**COMPUTER SCIENCE & ENGINEERING**

**IN BACHELOR OF TECHNOLOGY**

**JULY- AUGUST 2020**

**A REPORT OF THREE WEEKS INDUSTRIAL TRAINING**

**AT**

**WebTek Labs Pvt. Ltd**

**SUBMITTED IN PARTIAL FULFILLMENT OF THE**

**REQUIREMENT FOR THE AWARD OF**

**THE DEGREE OF**

**BACHELOR OF TECHNOLOGY**

**COMPUTER SCIENCE & ENGINEERING**



**DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING**

**CANDIDATE'S DECLARATION**

We hereby declare that we have undertaken industrial training at “WEBTEK LABS PVT. LTD.” during a period from 27th July to 16th August in partial fulfilment of requirements for the award of degree of B.Tech (COMPUTER SCIENCE & ENGINEERING) at **SHRI VAISHNAV VIDHYAPEETH VISHWAVIDYALAYA ,INDORE** The work which is being presented in the training report submitted to Department of COMPUTER SCIENCE & ENGINEERING at **SHRI VAISHNAV INSTITUTE OF TECHNOLOGY & SCIENCE, INDORE** is an authentic record of training work.

**STUDENTS NAME**

ALINA KHAN

SHEFALI PARMAR

PRAPTI PAGNIS

**ACKNOWLEDGEMENT**

It gives us great pleasure to acknowledge the guidance, assistance and support of Ms. Mousita Dhar in making the Project and this Project report successful, which has been structured under her valued suggestion.

She has helped us to accomplish the challenging task in a very short period of time.

Finally, we express the constant support of our friends, family and professors for inspiring us throughout and encouraging us.

ALINA KHAN

SHEFALI PARMAR

PRAPTI PAGNIS

SEMESTER:VII

CSE

**CERTIFICATE OF APPROVAL**

The project **“NETFLIX VISUALIZATION AND RECOMMENDATION SYSTEM**” made by the efforts of the team members is hereby approved as a creditable study for the Bachelor of Technology in COMPUTER SCIENCE & ENGINEERING and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

Ms. Mousita Dhar

**1.INTRODUCTION**

**1.1 PYTHON**

**About:**

* Python is a widely used general-purpose, high level programming language. It was created by **Guido van Rossum** in 1991 and further developed by the Python Software Foundation.
* It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.
* It is used for Web Development (server-side), Software Development, Mathematics, System Scripting.
* Python can be used on a server to create web applications.
* Python can be used alongside software to create workflows.
* Python can connect to database systems. It can also read and modify files.
* Python can be used to handle big data and perform complex mathematics.
* Python can be used for rapid prototyping, or for production-ready software development.

**Python versions:**

**Version1: Python1.0**

In January 1994, the first version of **Python 1.0** was released. This version 1 includes the major new features like the functional programming tools **filter**, **reduce**, **map**, and **lambda** etc.

**Version2: Python 2.0**

After Six and a half years later, Python 2.0 was introduced in October 2000. In this release, a full garbage collector, list comprehensions were included, and it also supports Unicode.

**Version3: Python 3.0**

Python then after 8 years, the next major release was made. This release was Python 3.0 also known as”Py3K” or “Python 3000”.

The major changes in Python 3.0 are:

* In this version, **Print** is a [Python function](https://intellipaat.com/blog/tutorial/python-tutorial/python-functions/)
* Instead of lists, in this version, we have **Views** and **iterators**
* In this version, we have more simplified rules for ordering comparisons. For example, we cannot sort a heterogeneous list, because each element of a [Python List](https://intellipaat.com/blog/tutorial/python-tutorial/python-lists/) must be comparable to other elements.
* In this python version, **int. long** is also an **int** as there is only one integer type.
* In this python version, when we divide two integers it resultant returns is a float instead of an integer. We can use **“//”** to have the “old” behavior.
* In this python version, Instead of **Unicode Vs. 8-bit** we have **Text Vs. Data**
* The one drawback of Python 3.0 is that it is not backward compatible with Python 2.x.

**Current Versions:**

**Python 3.8.0** is the newest major release of the Python programming language, and it contains many new features and optimizations.Some of them are as follows:

* Assignment expressions
* Positional-only arguments
* Python Initialization Configuration (improved embedding)
* Runtime audit hooks
* **continue** is now legal in **finally:** blocks

There are many other changes.

**Features :**

Some of the features of python are as follows:

* Easy to code
* Free and Open Source
* Object-Oriented Language
* GUI Programming Support:.
* High-Level Language
* Extensible feature
* Python is Portable language
* Python is Integrated language
* Interpreted Language:.
* Large Standard Library
* Dynamically Typed Language

**IDLE:**

IDLE (Integrated Development and Learning Environment) is an integrated development environment (IDE) for Python. IDLE can be used to execute a single statement just like Python Shell and also to create, modify and execute Python scripts. IDLE provides a fully-featured text editor to create Python scripts that includes features like syntax highlighting, autocompletion and smart indent. It also has a debugger with stepping and breakpoints features.

**1.2Anaconda**

Anaconda is an open-source distribution for python and R. It is used for datascience, machine learning, deep learning, etc. With the availability of more than 300 libraries for data science, it becomes fairly optimal for any programmer to work on anaconda for data science. It is easy to download and install, and it is supported on Linux, MacOS, and Windows. Conda makes it quick and easy to install ,run, and upgrade complex data science and machine learning environments like scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Service Machine Learning AMIs and Anaconda for Microsoft on Azure and Windows.

**1.3 Packages**

**1.3.1 NUMPY:**

NumPy is a module for Python. The name is an acronym for "Numeric Python" or "Numerical Python".  It is an extension module for Python, mostly written in C. This makes sure that the precompiled mathematical and numerical functions and functionalities of Numpy guarantee great execution speed. NumPy enriches the programming language Python with powerful data structures, implementing multi-dimensional arrays and matrices. These data structures guarantee efficient calculations with matrices and arrays. The implementation is even aiming at huge matrices and arrays, better know under the heading of "big data". Besides that the module supplies a large library of high-level mathematical functions to operate on these matrices and arrays.

NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

**1.3.2 PANDAS**

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

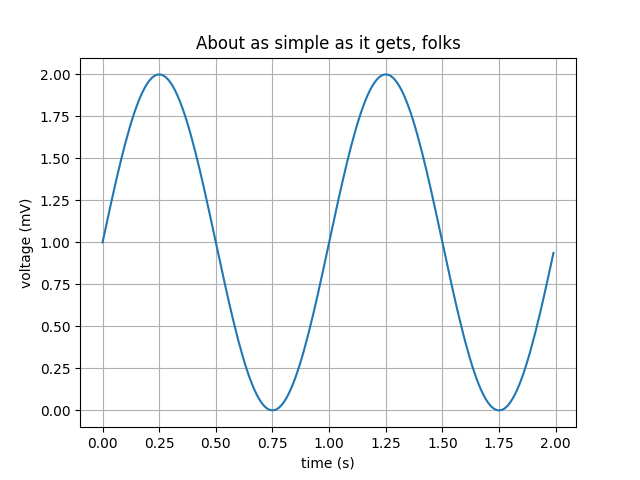
**1.3.3 MATPLOTLIB**

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

Matplotlib was originally written by John D. Hunter in 2003. The current stable version is 2.2.0 released in January 2018.

## Line Plot

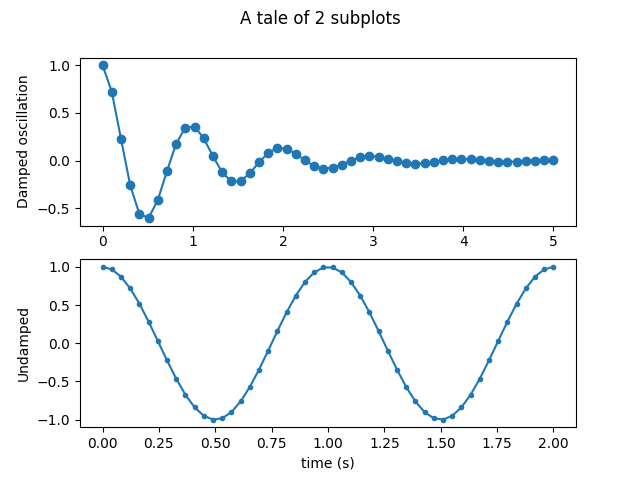
## Here's how to create a line plot with text labels using [plot()](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#matplotlib.pyplot.plot).



**Simple Plot**

## Multiple subplots in one figure

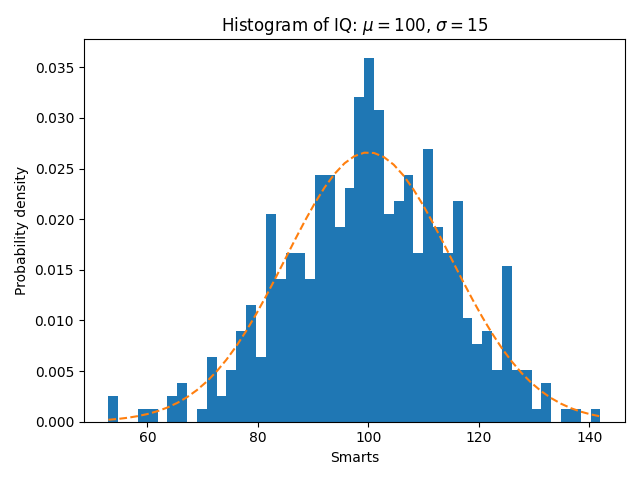
Multiple axes (i.e. subplots) are created with the [subplot()](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.subplot.html#matplotlib.pyplot.subplot) function:



**Subplot**

## Histograms

The [hist()](https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html#matplotlib.pyplot.hist) function automatically generates histograms and returns the bin counts or probabilities:



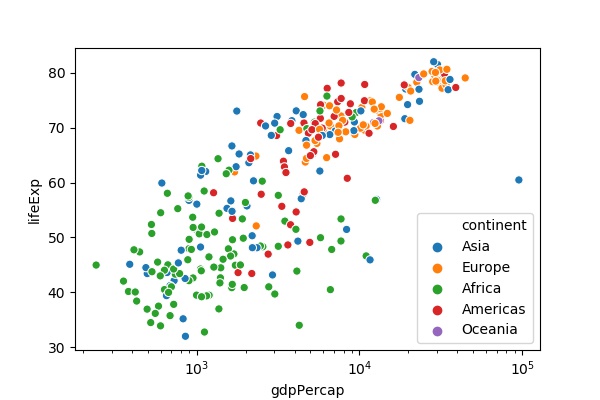
**Histogram**

**1.3.4 SEABORN**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from [pandas](https://www.geeksforgeeks.org/introduction-to-pandas-in-python/).  
Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

* **Scatterplot**

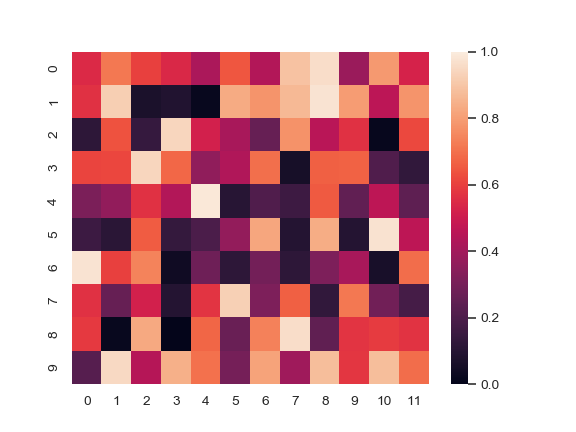
**Scatter Plot** represents the relationship between two continuous values, respectively. It depicts how one data variable gets affected by the other data variable in every fraction of the value of the data set.



SCATTER PLOT

* **Heatmap**

A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator’s requirement.



**1.3.5 SCIKIT-LEARN**

Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

* **NumPy**: Base n-dimensional array package
* **SciPy**: Fundamental library for scientific computing
* **Matplotlib**: Comprehensive 2D/3D plotting
* **IPython**: Enhanced interactive console
* **Sympy**: Symbolic mathematics
* **Pandas**: Data structures and analysis

Extensions or modules for SciPy care conventionally named [SciKits](http://scikits.appspot.com/scikits). As such, the module provides learning algorithms and is named scikit-learn.

**2.TRAINING WORK UNDERTAKEN**

**2.1 Collecting Data From Kaggle**

Kaggle, a subsidiary of [Google LLC](https://en.wikipedia.org/wiki/Google_LLC), is an online community of [data scientists](https://en.wikipedia.org/wiki/Data_science) and machine learning  practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education.

Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users.

**2.2 Data Science**

Data science is all about using data to solve problems. The problem could be decision making such as identifying which email is spam and which is not. So, the core job of a data scientist is to understand the data, extract useful information out of it and apply this in solving the problems.

When it comes to data science, we need some sort of programming language or tool, like Python. Although there are other tools for data science, like R and SAS. Python as a programming language has become very popular in recent times. It has been used in data science, IoT, AI, and other technologies, which has added to its popularity.

Python is used as a programming language for data science because it contains costly tools from a mathematical or statistical perspective. It is one of the significant reasons why data scientists around the world use Python. If you track the trends over the past few years, you will notice that Python has become the programming language of choice, particularly for data science.

There are several other reasons why Python is one of the most used programming languages for data science, including:

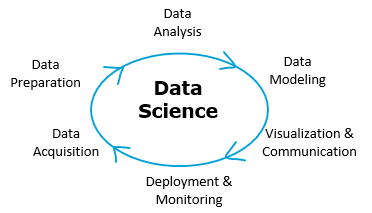
* **Speed** - Python is relatively faster than other programming languages
* **Availability** - There are a significant number of packages available that other users have developed, which can be reused
* **Design goal** - The syntax roles in Python are intuitive and easy to understand, thereby helping in building applications with a readable codebase.

To perform data analysis, you need to import specific libraries. Some examples include:

* **Pandas** - Used for structured data operations
* **NumPy** - A powerful library that helps you create n-dimensional arrays
* **SciPy** - Provides scientific capabilities, like linear algebra and Fourier transform
* **Matplotlib** - Primarily used for visualization purposes
* **Scikit-learn** - Used to perform all machine learning activities

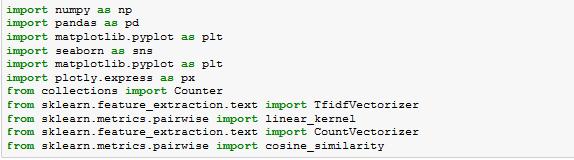
In addition to these, there are other libraries as well, like:

* **Networks & I graph**
* **TensorFlow**
* **BeautifulSoup**
* **OS**

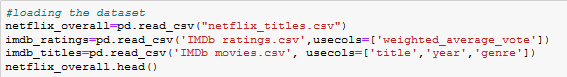


**2.2 SOURCE CODE & OUTPUT**

**1. IMPORT PACKAGES**

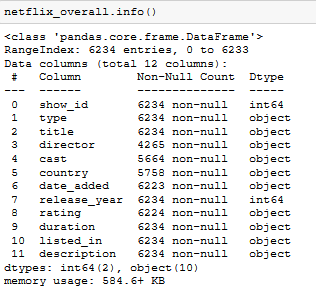
****

**2.LOADING DATASETS**

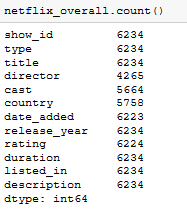


**3.DATA PREPROCESSING**

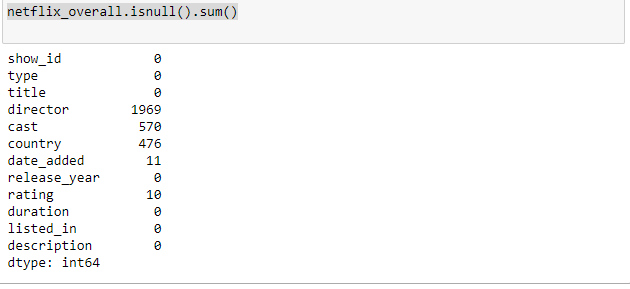
* Information of original dataset



* Counting non-NA cells for each column or row

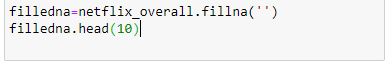


* Counting missing values for all columns

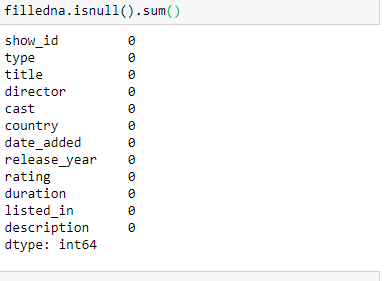


By using **isnull().sum()** we found that the **director** column has more than 30% values missing

* For the missing values we filled them with **‘ ‘**



After this we can again check for missing values



Now there are no missing values

**4.RECOMMENDATION SYSTEM**

For the recommendation part we are using **Content based filtering , TF-IDF (Term Frequency –Inverse Document Frequency), Cosine Similarity**

* **Content based filtering:**

In this system, the objects are mainly defined by their associated features. A content-based recommender learns a profile of the new user’s interests based on the features present, in objects the user has rated. It’s basically a keyword specific recommender system here keywords are used to describe the items. Thus, in a content-based recommender system the algorithms used are such that it recommends users similar items that the user has liked in the past or is examining currently.

In our project we apllied content based filtering on following factors:

**Title**

**Cast**

**Director**

**Listed in**

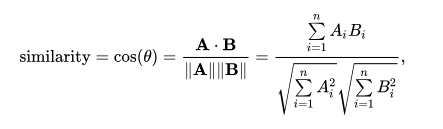
**Plot**

* **TF-IDF (Term Frequency –Inverse Document Frequency):**

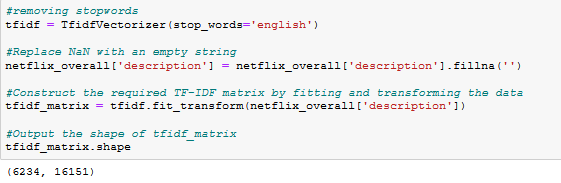
Score is the frequency of a word occurring in a document, down-weighted by the number of documents in which it occurs. This is done to reduce the importance of words that occur frequently in plot overviews and therefore, their significance in computing the final similarity score.

* **Cosine Similarity:**

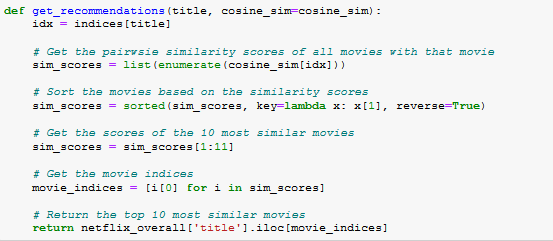
Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

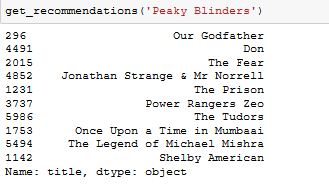


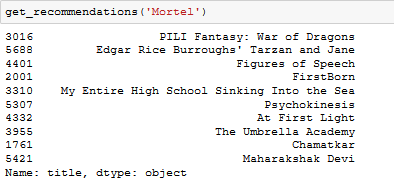
* **Recommendation System (Based on Content)**

****

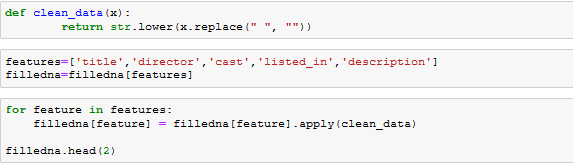
**C:\Users\Ankit\Prapti training\SCREENSHOTS\recommendation_system_part2.PNG**

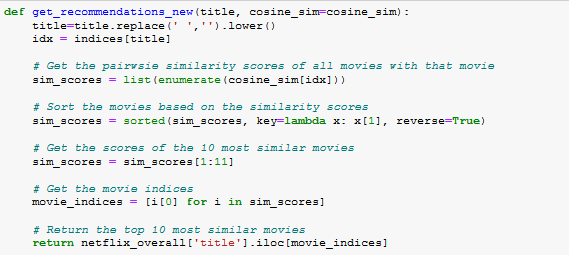
****

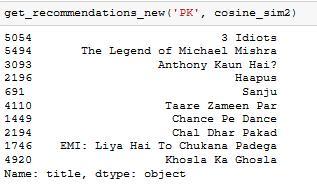
****

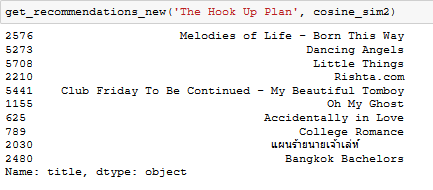
****

* **Recommendation System (Based on Multiple Matrics)**

****

****

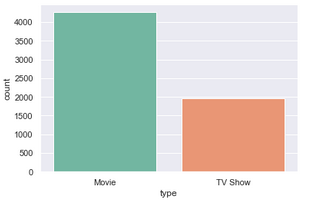
****

****

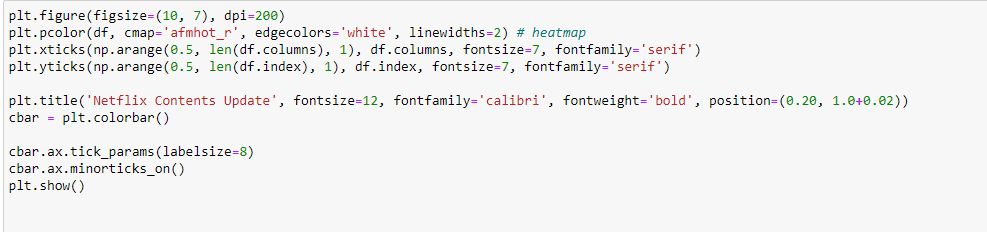
**5.DATA VISUALIZATION**

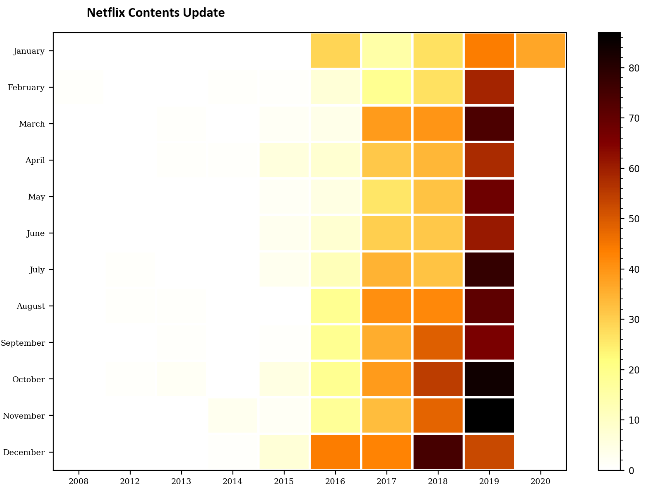
* **TV Show Vs Movies Graph**

C:\Users\Ankit\Prapti training\SCREENSHOTS\TV Shows vs Movies code.PNG



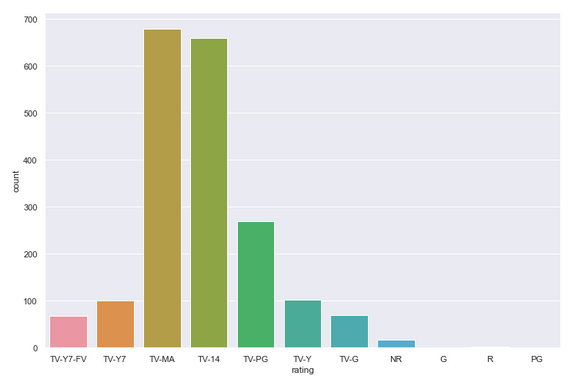
* **Finding out which month is best for some release, according to previous data**





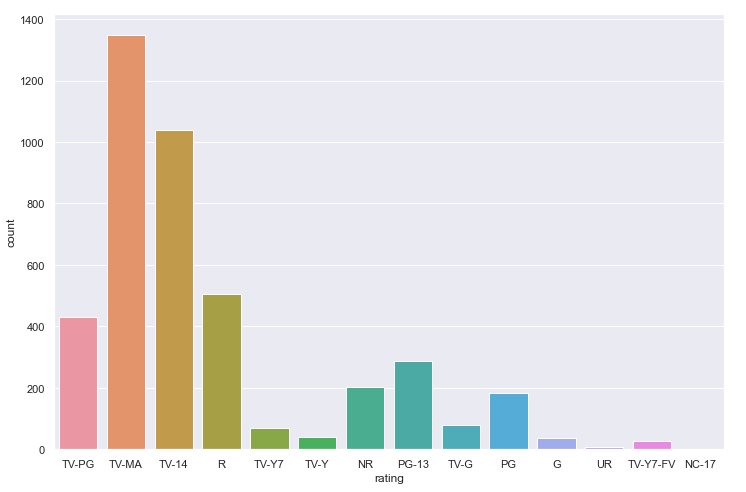
* **TV Show Rating analysis**

**C:\Users\Ankit\Prapti training\SCREENSHOTS\TV Show Rating analysis code.PNG**

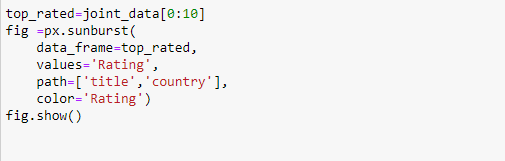
****

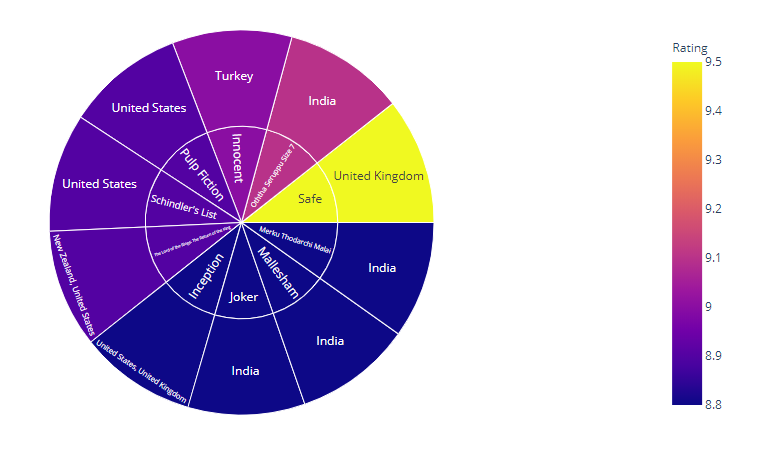
* **Movie Rating Analysis**

**C:\Users\Ankit\Prapti training\SCREENSHOTS\Movies rating analysis code.PNG**

****

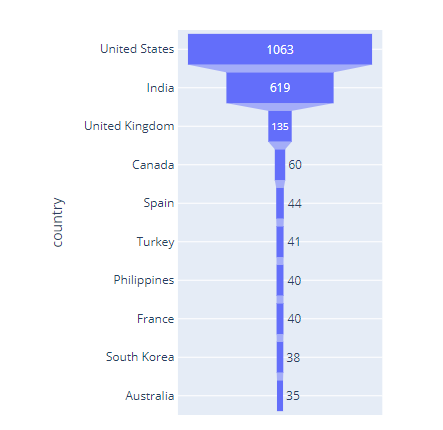
* **Top 10 rated Movies**

****

****

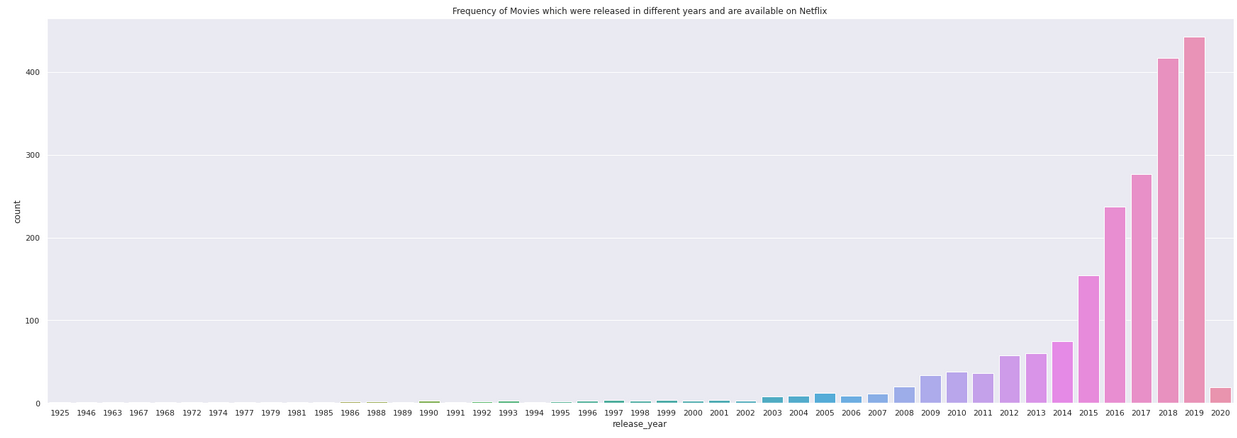
* **Countries with highest content**

****

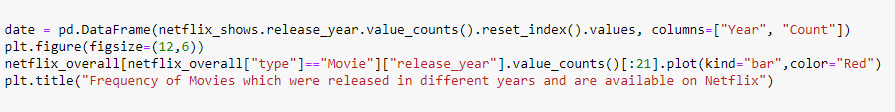
****

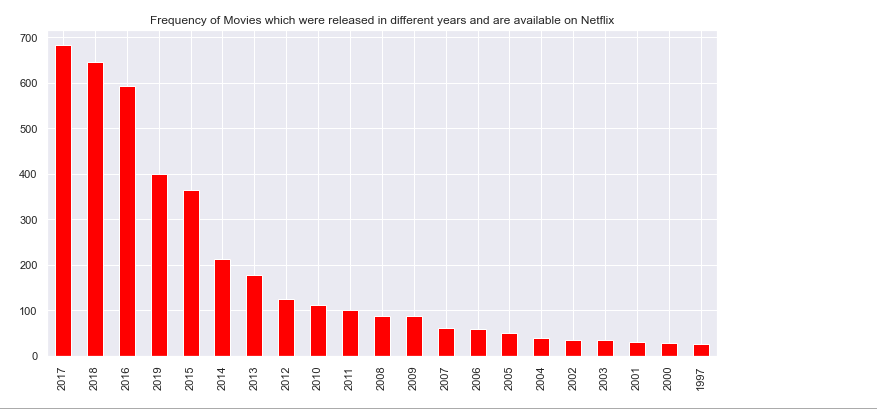
* **Year wise analysis**
* **TV Shows**

**C:\Users\Ankit\Prapti training\SCREENSHOTS\Year wise analysis TV Shows code.PNG**

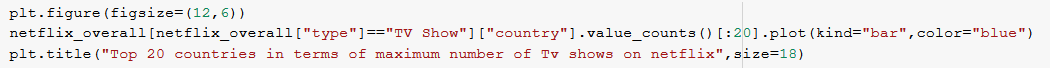
****

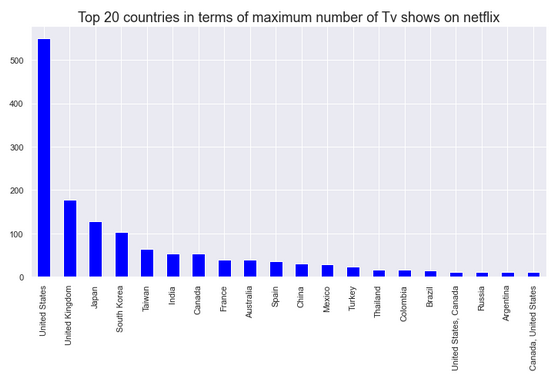
* **Movies**

****

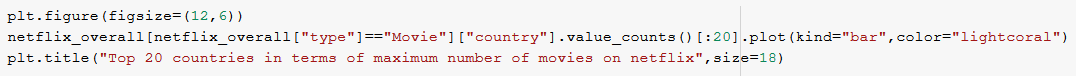
****

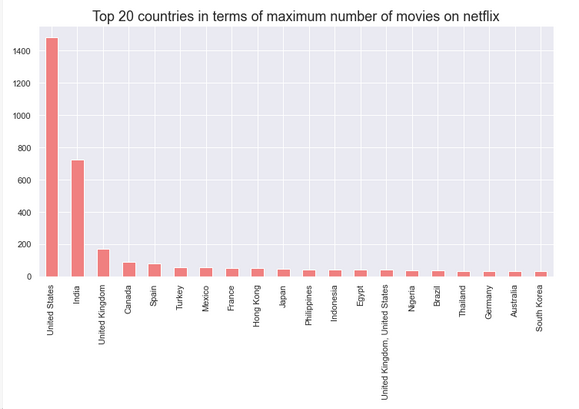
* **Top 20 content creating countries( for TV shows)**

****

****

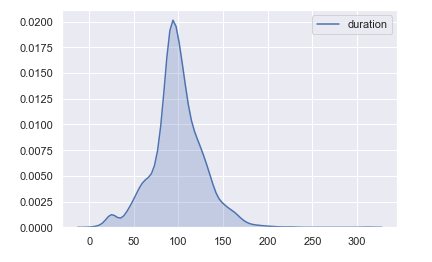
* **Top 20 content creating countries( for Movies)**

****

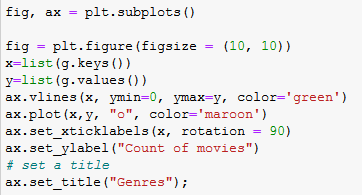
****

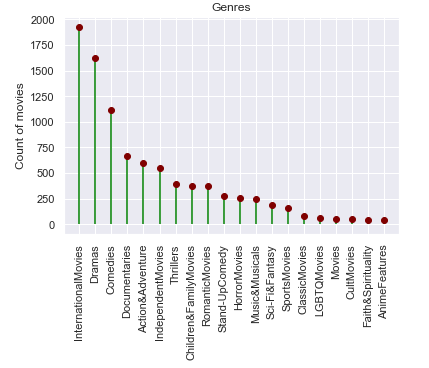
* **Analysis of Duration of Movies**

****

****

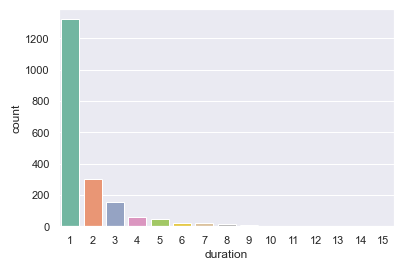
* **Plot graph according to Genre**

****

****

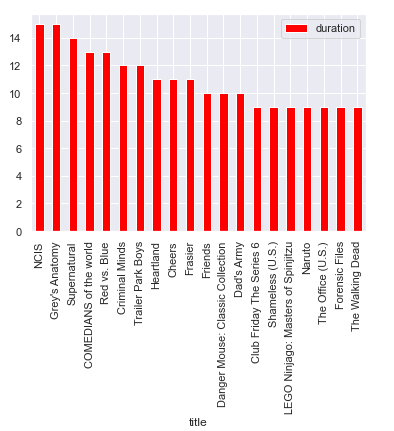
* **TV shows with seasons count**

**C:\Users\Ankit\Prapti training\SCREENSHOTS\analysis of no. of seasons code.PNG**

****

* **Top 20 TV show with highest number of seasons**

**C:\Users\Ankit\Prapti training\SCREENSHOTS\Tv shows with highest no. of seasons code.PNG**

****

**6.CONCLUSION**

The main goal of this project is to recommend Tv shows and Movies to the users.

Another goal was to highlight the rating of Tv shows and Movies

We have done ratings on different basis i.e. Tv shows with seasons count , year wise analysis etc.

Recommender systems are among the most popular applications of data science. They are used to predict the "rating" or "preference" that a user would give to an item.

This project involves Content based technique **Term Frequency Inverse Document Frequency (TF/IDF),Cosine similarity.**

They have the ability to recommend new items even if there are no ratings provided by users. So even if the database does not contain user preferences, recommendation accuracy is not affected. Also, if the user preferences change, it has the capacity to adjust its recommendations in a short span of time.

**REFERENCES**

* <https://www.kaggle.com/>
* [www.geeksforgeeks.org](http://www.geeksforgeeks.org)
* [www.python.org](http://www.python.org)
* <https://realpython.com/>
* <https://stackoverflow.com/>