Project Report
on

NETFLIX DATA ANALYSIS WITH PYTHON

TECHNICAL SKILLING -2

By

N. Sowgna (2010030344)

Tahseen Begum (2010030168)

E. Pravallika (2010030046)

Keerthana (2010030445)

under the supervision of

Chanda Raj Kumar Rao

Assistant Professor

Department of Computer Science and Engineering

K L University Hyderabad,

Aziz Nagar, Moinabad Road, Hyderabad – 500 075, Telangana, India.

March, 2022

Declaration

The Project Report entitled "NETFLIX DATA ANALYSIS WITH PYTHON" is a record of bonafide work of 2010030344(N.Sowgna), 2010030046 (E.Pravallika), 2010030168(Tahseen Begum) , 2010030445(Keerthana Pulugam), submitted as a requirement for the completion of the course **TECHNICAL SKILLING -2** in the Department of Computer Science and Engineering to the K L University, Hyderabad. The results embodied in this report have not been copied from any other Departments/University/Institute.

Signature of the Students

Certificate

This is to certify that the Project Report entitled "NETFLIX DATA ANALYSIS WITH PYTHON" is being submitted by 2010030344(N.Sowgna),2010030046(E.Pravallika), 2010030168 (Tahseen Begum) , 2010030445 (Keerthana Pulugam) ,as a requirement for the completion of the course in the **TECHNICAL SKILLING -2** Department of Computer Science and Engineering, K L University, Hyderabad is a record of bonafide work carried out under our guidance and supervision. The results embodied in this report have not been copied from any other departments/ University/Institute.

Signature of the Supervisor

Name and Designation

Signature of the HOD

Signature of the Examiner

ACKNOWLEDGEMENT

First and foremost, we thank the lord almighty for all his grace & mercy showered upon us, for completing this project successfully.

We take grateful opportunity to thank our beloved Founder and Chairman who has given constant encouragement during our course and motivated us to do this project. We are grateful to our Principal **Dr. L. Koteswara Rao** who has been constantly bearing the torch for all the curricular activities undertaken by us.

We pay our grateful acknowledgement & sincere thanks to our Head of the Department **Dr. Chiranjeevi Manike** for his exemplary guidance, monitoring and constant encouragement throughout the course of the project. We thank **Chanda Raj Kumar Rao** of our department who has supported throughout this project holding a position of supervisor.

We whole heartedly thank all the teaching and non-teaching staff of our department without whom we won't have made this project a reality. We would like to extend our sincere thanks especially to our parents, our family members and friends who have supported us to make this project a grand success.

1.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 article, I'm going to introduce you to a data science project on Netflix data analysis with Python.

We can analyze 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. 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 article, I'm going to introduce you to a data science project on Netflix data analysis with Python.

TABLE OF CONTENTS

2.INTRODUCTION

I'll take a look at some very important models of Netflix data to understand what's best for their business. Some of the most important tasks that we can analyze from Netflix data are:

Understand what content is available
Understand the similarities between the content
Understand the network between actors and directors
What exactly Netflix is focusing on
Sentiment analysis of content available on Netflix.

3.LITERATURE SURVEY

Author	Title	Publishing	Dataset & Techniques	Pros
Rajeswari Nakka Dr.G.V.S.N.R.V.Prasadand R.Kiran Kumar	Offering Recommendations on Netflix dataset by Associations among Users as Trust Metric	2021	By this model several batches of data points are sampled among total dataset points. The results are well analysed over error rate where the proposed technique tends to have the reduced error value compared to existing Collaborative Filtering technique. In future the work can be extended on larger data points and address issues like memory issues, and utilize optimized matrix factorization technique to move a step further.	The model is evaluated using an evaluation metric Mean Squared Error (MSE) value. The performance of the proposed algorithm on the Netflix dataset was compared with existing algorithm
Vadloori, Karthik Babu, and Shriya Madhavi Sanghishetty.	Exploratory and Sentiment Analysis of Netflix Data	2021	An Open-Source Data Set obtained from Kaggle – that was wrangled and exercised to derive maximum insights using EDA – Exploratory Data Analysis and Sentiment Analysis after the amalgamation of two additional sets – Geographical Latitudes & Longitudes and Netflix Title Critics/Reviews Data Set. The project is made using different utility analytical tools present in Python Library of versatile packages	Introduces systematic and insightful usage of methods for Exploratory Data Analysis & Sentiment Analysis by utilizing various packages concerned.
Eva-Patricia Fernandez-Manzano, Elena Neira, Judith Clares-Gavilan	Data management in the audio visual industry: Netflix as a case study	July-August 2020	Big data has become an enormous asset for on-demand content distribution services, helping information supply and decision-making, regarding both the content of the database and subscribers to the database. In this article, we describe and define big data and data management in a media company devoted to on-demand audio visual content distribution: Netflix.	The model is evaluated using an evaluation metric Mean Squared Error (MSE) value. The performance of the proposed algorithm on the Netflix dataset

4.HARDWARE & SOFTWARE REQUIREMENTS

1)WINDOWS 11 FOR X64-BASED SYSTEM:

Ensure the PC you want to install Windows 11: Has a 64-bit CPU: Windows 11 can only run on 64-bit CPUs. To see if your PC has one, go to Settings > System > About, or search "System Information" in Windows and look under "System Type." The Windows 11 ISO is only available for devices with x64 processors.

2) BRACKETS:

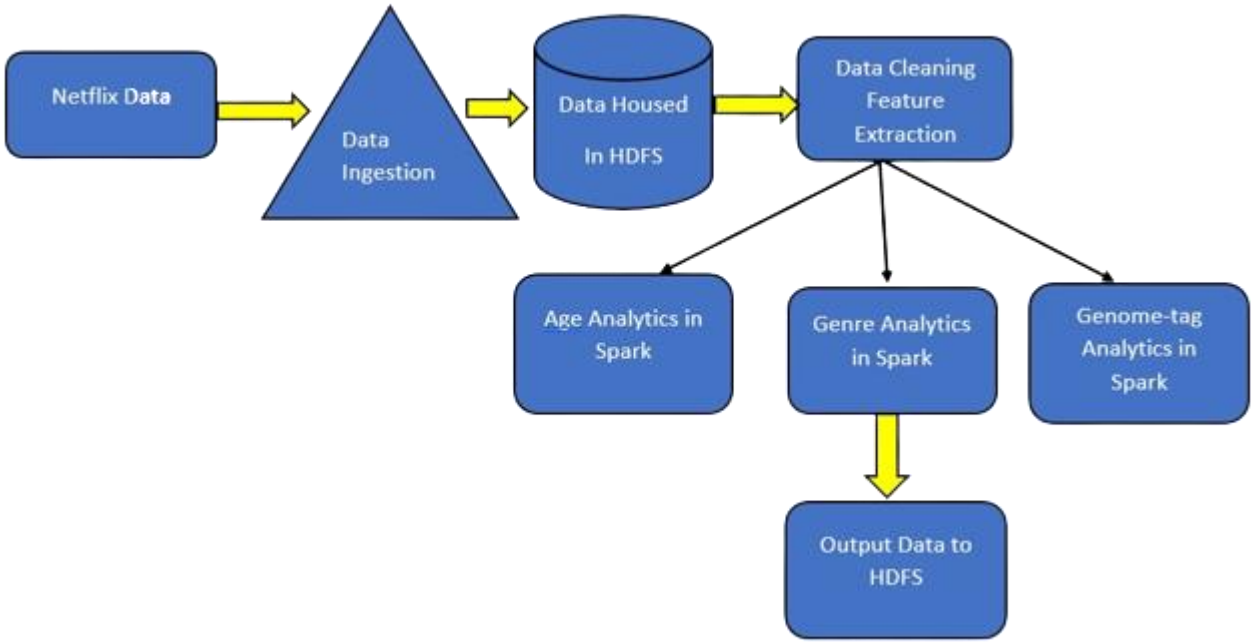
Brackets is a source code editor with a primary focus on web development. Created by Adobe Inc., it is free and open-source software licensed under the MIT License, and is currently maintained on GitHub by open-source developers. It is written in JavaScript, HTML and CSS. Brackets is cross-platform, available for macOS, Windows, and most Linux distributions. The main purpose of Brackets is its live HTML, CSS and JavaScript editing functionality.

3) JUPITER NOTEBOOK:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

5.FUNCTIONAL & NON-FUNCTIONAL REQUIREMENTS

5.1 FLOW CHART:



6.PROPOSED SYSTEM

- Importing Libraries
- Loading the dataset
- Data Cleaning:
- Deleting redundant columns.
- Dropping duplicates.
- Cleaning individual columns.
- Remove the NaN values from the dataset
- Some Transformations
- Data Visualization: Using plots to find relations between the features.
- Type: Movie and TV Shows
- Rating
- Relation between Type and Rating
- Word Cloud
- Country
- Cast
- Director
- Category

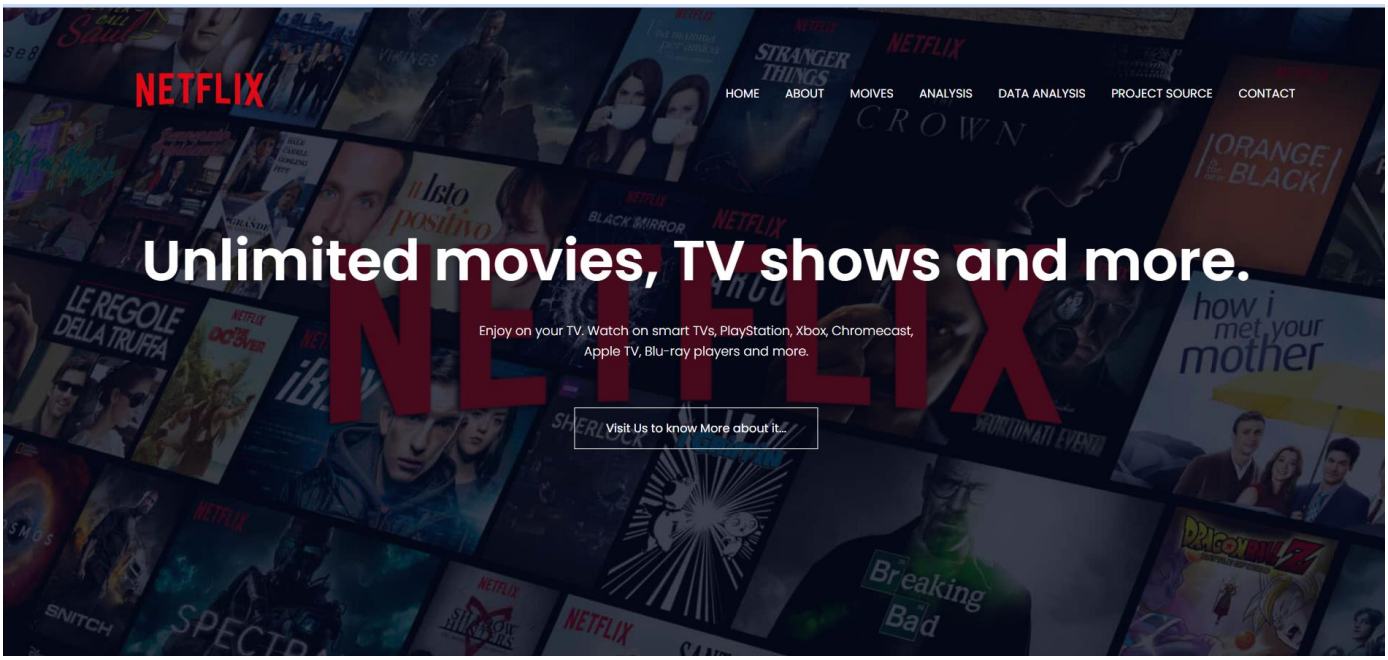
7.IMPLEMENTATION

Step-1:
```
!pip install matplotlib
```
Step-2:
```
#importing the dataset
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
```
Step-3:
```
#to import csv file
netflix_data = pd.read_csv(r"C:\Users\Sowgn\OneDrive\Desktop\pfsd\netflix_titles.csv")
```
Step-4:
```
netflix_data
```
Step-5:
```
#head function
#to show top-5 records of the dataset
netflix_data.head()
```
Step-6:
```
#tail function
#to show bottom-5 records of dataset
netflix_data.tail()
```
Step-7:
```
#shape
#to show no of rows and colums
netflix_data.shape
```
Step-8:
```
#size
#to show no of total values(elements) in the dataset
netflix_data.size
```
Step-9:
```
netflix_data.columns
```
Step-10:
```
netflix_data.dtypes
```

Step-11:
```
netflix_data.info()
```
Step-12:
```
netflix_data.value_counts()
```
Step-13:
```
netflix_data.nunique()
```
Step-14:
```
netflix_data[netflix_data.duplicated()]
```
Step-15:
```
netflix_data.drop_duplicates()
```
Step-16:

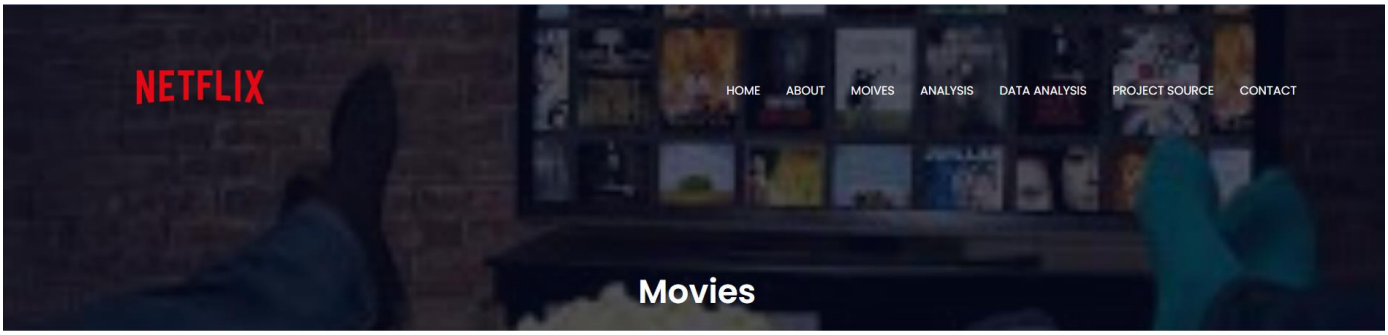
netflix_data.isnull()
Step-17:
netflix_data.isnull().sum()
Step-18:
!pip install seaborn
Step-19:
import seaborn as sns
Step-20:
sns.heatmap(netflix_data.isnull())
Step-21:
netflix_data[netflix_data['title'].isin(['Blood & Water'])]
Step-22:
netflix_data['Date_N']=pd.to_datetime(netflix_data['date_added'])
Step-23:
netflix_data.head()
Step-24:
netflix_data['Date_N'].dt.year.value_counts()
Step-25:
netflix_data['Date_N'].dt.year.value_counts().plot(kind='bar')
Step-26:
netflix_data.groupby('type').type.count()
Step-27:
sns.countplot(netflix_data['type'])
Step-28:
netflix_data['year']=netflix_data['Date_N'].dt.year
netflix_data.head(2)
Step-29:
netflix_data[(netflix_data['type']=='Movie')&(netflix_data['year']==2022)]

8.RESULTS DISCUSSION



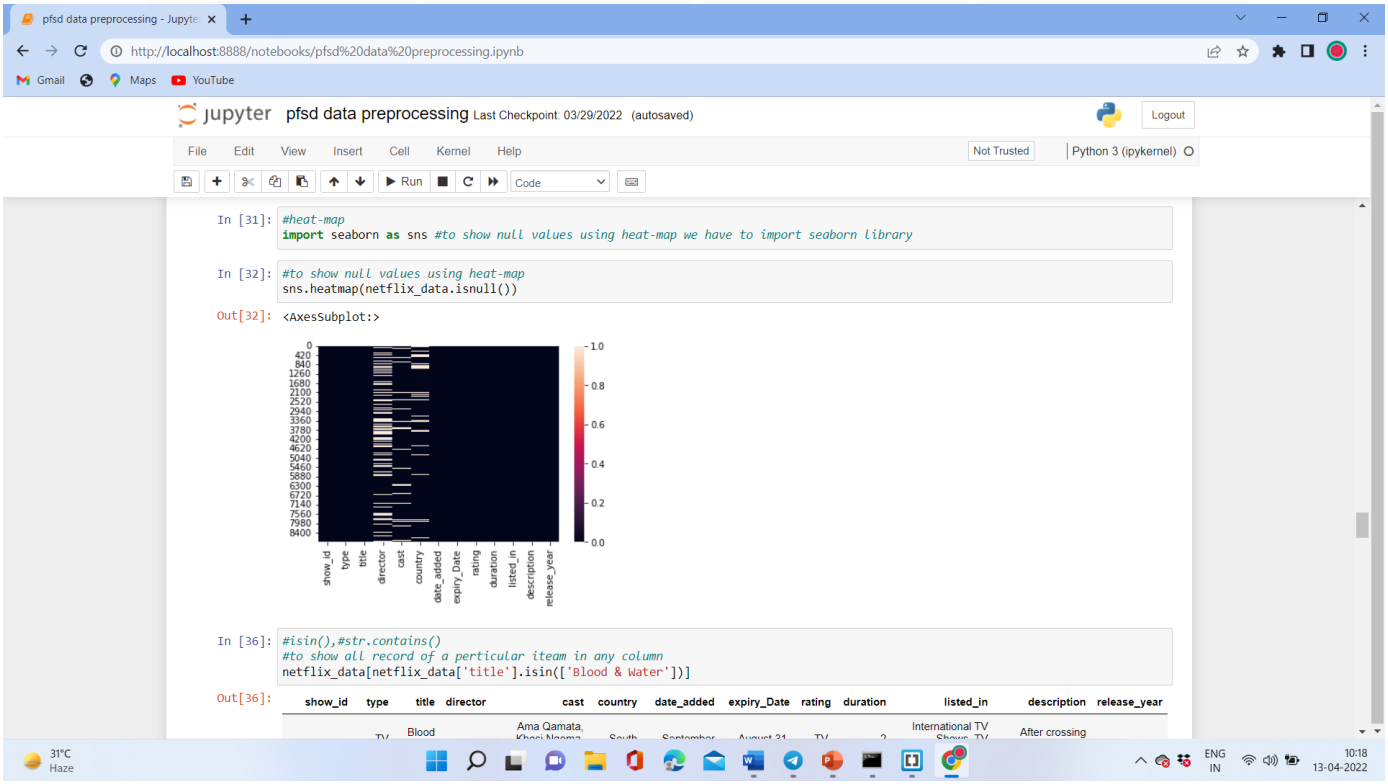
Unstoppables

We are students from K L University, Hyderabad. We are BTech students from CSE branch (Honors). We team members Successfully completed our PROJECT.



Enjoy on your TV.

Watch on smart TVs, PlayStation, Xbox, Chromecast, Apple TV, Blu-ray players and more.



pfsd data preprocessing - Jupyte

http://localhost:8888/notebooks/pfsd%20data%20preprocessing.ipynb

Gmail Maps YouTube

jupyter **pfsd data preprocessing** Last Checkpoint: 03/29/2022 (autosaved)

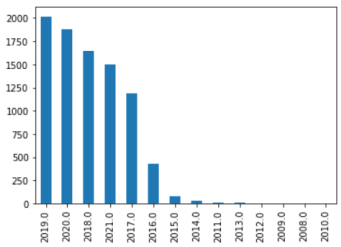
Logout

File Edit View Insert Cell Kernel Help | Not Trusted | Python 3 (ipykernel)

Run Code

In [41]:
```
#bar graph
netflix_data['Date_N'].dt.year.value_counts().plot(kind='bar')
```

Out[41]: <AxesSubplot:>



In [44]:
```
#groupby(),to group all unique items of a column and show their count
netflix_data.groupby('type').type.count()
```

Out[44]:
```
type
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [45]:
```
sns.countplot(netflix_data['type'])
```

```
C:\Users\Sowgn\AppData\Roaming\Python\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments with
```

31°C Haze

ENG IN 10:18 13-04-2022

pfsd data preprocessing - Jupyte

http://localhost:8888/notebooks/pfsd%20data%20preprocessing.ipynb

Gmail Maps YouTube

jupyter **pfsd data preprocessing** Last Checkpoint: 03/29/2022 (autosaved)

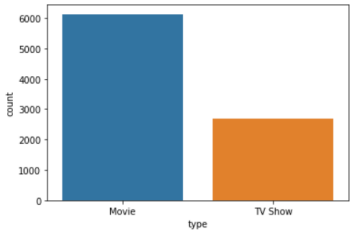
Logout

File Edit View Insert Cell Kernel Help | Not Trusted | Python 3 (ipykernel)

Run Code

```
C:\Users\Sowgn\AppData\Roaming\Python\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[45]: <AxesSubplot:xlabel='type', ylabel='count'>



In [46]:
```
#creating new column
netflix_data['year']=netflix_data['Date_N'].dt.year
```

In [47]:
```
netflix_data.head(2)
```

Out[47]:

	show_id	type	title	director	cast	country	date_added	expiry_Date	rating	duration	listed_in	description	release_year	Date_N	year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	August 31, 2021	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	2020	2021-09-25	2021.0

31°C Haze

ENG IN 10:18 13-04-2022

9.CONCLUSION AND FUTURE WORK

We have drawn many interesting inferences from the dataset Netflix titles

- The most content type on Netflix is movies.
- The country by the amount of the produces content is the United States,
- The most popular director on Netflix , with the most titles, is Jan Suter.
- International Movies is a genre that is mostly in Netflix.
- largest count of Netflix content is made with a “TV-14” rating.
- The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai.
- The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.

It's clear that Netflix has grown over the years. We can see it from the data that the company took certain approaches in their marketing strategy to break into new markets around the world. some future projects to work on : netflix recommendation system and text analysis.

10.REFERENCES

- https://www.researchgate.net/publication/345710976_Offering_Recommendations_on_Netflix_dataset_by_Associations_among_Users_as_Trust_Metric
- https://www.researchgate.net/publication/354719521_Exploratory_and_Sentiment_Analysis_of_Netflix_Data
- https://www.researchgate.net/publication/305741976_Data_management_in_audiovisual_business_Netflix_as_a_case_study
- https://www.youtube.com/watch?v=b7Kd0fLwgO4
- https://www.kaggle.com/shivamb/netflix-shows