SURESH GYAN VIHAR UNIVERSITY

Accredited by NAAC with 'A' Grade

AN
INTERNSHIP REPORT
ON

"PRACTICAL TRAINING SEMINAR"

In partial fulfillment
For the award of Degree of
"Bachelor of Technology in CSE with AI-ML"

Submitted to:
Ms. Pooja Varshney
Assistant Professor
CEIT-DEPARTMENT

Submitted by:
Shyam Bihari Kumar (99695)
B. Tech CSE – Vth Sem

Department of Computer Engineering & Information Technology

STUDENT DECLARATION

I hereby declare that the project work entitled **"Data Science & Machine Learning Using Python"** submitted to the SGVU Jaipur, is a record of an original work done by me under the guidance of **Ms. Pooja Varshney**, Assistance Professor, Dept. of Computer Engineering and Information Technology, Gyan Vihar School of Engineering and Technology, SGVU.

This project work is submitted in the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering. This result embodied in this project have not been submitted to any other University or Institute for the award of any degree or diploma.

Student's Sign

Submitted To:

Ms. Pooja Varshney

(Assistant Professor)

CERTIFICATE

This is to certify that the people report entitled “Data Science & Machine Learning Using Python” Submitted to Suresh Gyan Vihar University, Jaipur in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology, is an authentic and original work carried out by **Shyam Bihari Kumar (99695)** under my guidance.

The matter embodied in this project is genuine work done by the student and has not been submitted whether to this university or to any university for the fulfillment of the requirement of any course of study.

Ms. Pooja Varshney
Assistant Professor

Dr. Sohit Agarwal
HOD, CEIT

ACKNOWLEDGEMENT

I would like to express my profound gratitude to Ms. Pooja Varshney of CEIT department, and Dr. Sohit Agarwal of Suresh Gyan Vihar University for their contributions to the completion of my project titled **"Data Science & Machine Learning Using Python".**

I would like to express my special thanks to our mentor **Ms. Pooja Varshney** for her time and efforts she provided throughout the year. Her useful advice and suggestions were really helpful to me during the project's completion. In this aspect, I am eternally grateful to her.

I would like to acknowledge that this project was completed entirely by me and not by someone else.

Student Name:

Shyam Bihari Kumar

SID No.: 99695

Sem: 5

ABSTRACT

The objective of this summer internship report is to generally present an overview of data science with machine learning techniques currently in use or taken into consideration by data scientists worldwide.

Aim of this internship is first to learn about data science, big data visualization, machine learning and deep learning, implement the following tools: anaconda, jupyter notebook, spider and the libraries following: numpy, pandas, matplotlib, seaborn, portly, scikit-learn by describing the platforms used by data scientists like kaggle and finally integrating the machine learning into e-government platform.
Since these terms mentioned above generally used in others place wrongly and create confusion in people's mind, each of them explained section by section. During this internship, more than one projects created in the fields of big data visualization and machine learning which aim to improve e-government platform by using new technologies. As a result, I've combined my old software engineering skills with this new data science experience and I am able to write new kernels and join Kaggle's machine learning competitions and forum discussions which are very useful for my career as a data scientist.

TABLE OF CONTENT

Introduction

1.1Background and Subjects Declarations

During my internship, I got a chance to work in the PHN Technology to know about how a software company uses big data in applications used by various public institutions, so the Company which I was working on naturally dealing with massive volume of data that concerns data science.

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract value from data.

Data scientists combine a range of skills—including statistics, computer science, and business knowledge, machine learning, deep learning ——to analyze data collected from the web, smartphones, customers, sensors, and other source. These data sets are so voluminous that traditional data processing software just can't manage them, So big data is becoming one of the most important technology trends that has the potential for dramatically changing the way organization's use information to enhance the customer experience and transform their models. For example,

- Big Data visualization is the graphical representation of information and data.

  By using visual elements like charts, graphs, and maps, data

  visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

This mass of data is useless unless we analyze it and find the patterns hidden within.

- Machine Learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. We'll classify it supervised and unsupervised learning and we'll implement

Data Science and Machine Learning with Python is an invitation to unlock the potential of data and the power of Python. It's a passport to a world where you can extract knowledge, make informed decisions, and craft intelligent solutions. So, let's embark on this exciting adventure and master the art of data science and machine learning with Python

In the modern age of information, data has become one of the most valuable assets. The ability to collect, analyse, and derive insights from data is a superpower that drives innovation and decision-making in almost every industry. Data Science and Machine Learning with Python is a journey into this dynamic and transformative field, where Python serves as your trusty tool to unlock the potential hidden within vast datasets.

Data Science Unveiled: Data science is the art of transforming raw data into meaningful insights. It's the process of asking the right questions, collecting and preparing data, exploring patterns and trends, and building predictive models. This field sits at the intersection of statistics, computer science, and domain expertise, and it's poised to reshape how we perceive the world.

The Power of Python: Python is your gateway to the data-driven world. Known for its simplicity and versatility, Python has become the go-to language for data science and machine learning. Its vast ecosystem of libraries and tools, such as NumPy, Pandas, Matplotlib, and Scikit-Learn, empowers data scientists and machine learning practitioners to tackle complex problems with ease.

.

About The Organization

PHN Technology is a private company in India that focuses on designing niche offerings for its clientele. Here are some key details about the company:

PHN Technology Pvt.Ltd. are extremely committed to designing solutions for our clients that help them meet their business needs in a time bound manner.

PHN Technology Pvt.Ltd. have a customer base which ranges close to 400 million customers that has been built over a period of the last decade.

PHN Technology Pvt.Ltd. are the right choice for organizations that need to scale their businesses by leveraging our already existing customer base to meet their business outcomes.

PHN Technology Pvt.Ltd. have ability to impact your P&L positively by keeping marketing costs for you minimalistic as they have the potential to ensure reach to their large customer base and ensure implementation and adoption of your products.

PHN Technology Pvt.Ltd. boast about their ability of ensuring their business needs are met and tied to their compensation model.

PHN Technology private limited is a Consultancy service based company which provide many software related service such as :

Business Solution: It is our fundamental belief that every offering we design for our clients' needs to be right priced while maintaining quality standards.

Robotics: We have programs designed for every grade and age level which is offered to students both as a regular program and as a post school activity.

Training & Education: We believe that for learners to be job ready we need to train them based on the core requirements of the company they would want to work.

Mobile App Development: We provide custom app development services for Android, iOS, Artificial Intelligence, Blockchain, Gaming & AR/VR.

Website Development: We assist our customers with developing their online website by utilizing proficient and creative website designers.

My Role and Experience

Data Collection and Cleaning:

- Collecting raw data from various sources.
- Cleaning and preprocessing data to ensure its quality.

Exploratory Data Analysis (EDA):

- Conducting EDA to understand the dataset.
- Identifying patterns, trends, and outliers.

Coding and Scripting:

- Writing and maintaining code in languages like Python or R.

Feature Engineering:

- Creating new features from existing data to improve model performance.

Data Visualization:

- Creating visualizations to communicate insights from data.

Model Development:

- Assisting in building machine learning models.
- Collaborating with the team to select appropriate algorithms.

Model Evaluation:

- Assessing model performance through metrics like accuracy, precision, and recall.
- Fine-tuning models based on evaluation results.

Learning Opportunities:

- Taking advantage of mentorship and learning from experienced data scientists.

Networking:

- Building professional connections within the organization.

Final Project:

-Work on a real project that contributes to the company's goals.

3.1 Phases of internship

Internship Domain: Machine learning & Data Science using Python

Tenure: 2 months

Internship Start Date- 06 April 2023

Internship End Date - 06 June 2023

Stipend: Unpaid

My role in this internship –As an Intern

Mode: Remote(WFH)

Data Science

4.1 Definitions and Declarations about Data Science

Data science is the future of Artificial Intelligence. Therefore, it is very important to understand what is Data Science and how can it add value to your business.
Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using the simple tools like SQL, Oracle, etc. Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured .This data is generated from different sources like financial logs, text files, multimedia forms, sensors and instruments. Simple tools mentioned above are not capable of processing this huge volume and variety of data. This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it and then act upon.

4.2 Tools and Platforms for Data Scientists

4.2.1 Data Collection Tools :

Collecting quality data that can be transformed into rich analysis is the starting point every data strategy. The right data collection tools can reduce errors and duplicates, ensure greater accuracy, and preserve the integrity of data coming from all sources.
E.g. :GoSpotCheck, IBM Datacap, Mozenda, Octoparse, etc.

4.2.2 Data Analysis Tools:

Finding meaning in and extracting value from your data is the core of all data analysis tools that enable you to easily understand and derive real meaning from your data help you male right business decisions that impact revenue and competitiveness.
Alteryx, Domino Data Lab , KNIME Analytics Platform, etc.

4.2.3 Data Warehousing Tools:

Data warehouses function as repositories for data that's been combined and integrated from multiple, disparate sources and then standardized for ease of use. E.g. Amazon RedShift, Google BigQuery, Microsoft Azure, MySQL, etc.

4.2.4 Data Visualization Tools:

Visual analytics tools identify patterns and trends in your data and help end users Understand and digest complex concepts.

Google Colab , Jupyter , Microsoft Excel ,SAS, etc.

4.2.5 Data Scientists Home: Kaggle:

Kaggle is an online community of data scientists and machine learners, owned by Google LLC. 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

How Data Science is Transforming Business

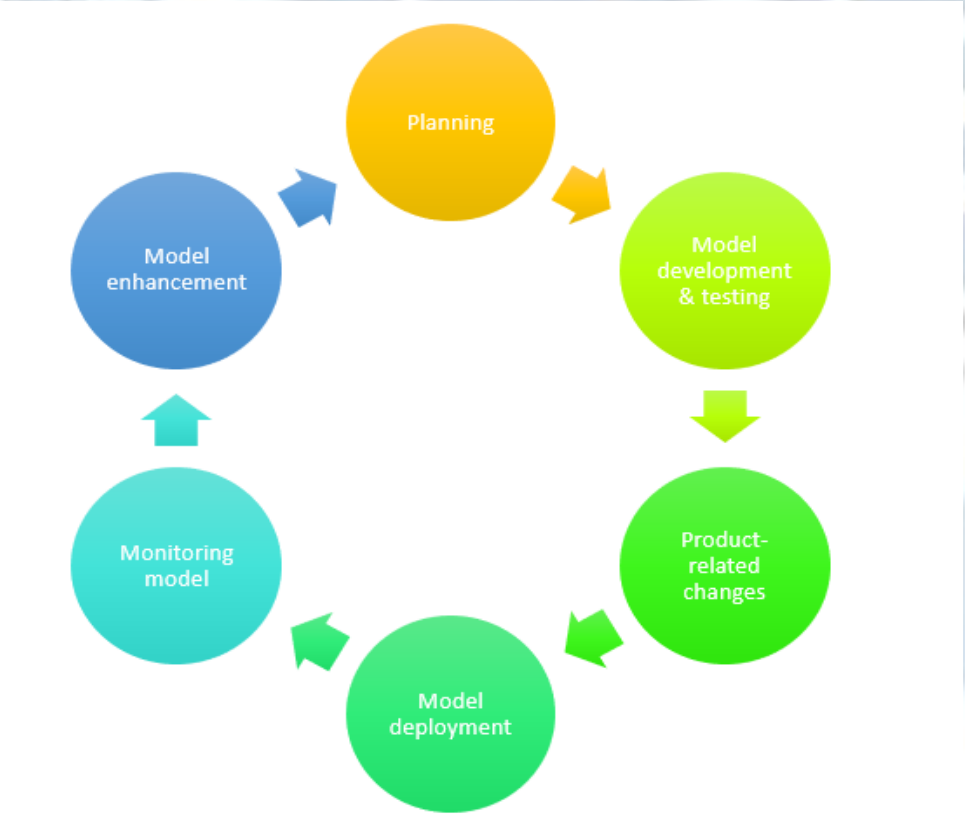
Organizations are using data science teams to turn data into a competitive advantage by refining products and services. For example, companies analyze data collected from call centers to identify customers who are likely to churn, so marketing can take action to retain them. Logistics companies analyze traffic patterns, weather conditions, and other factors to improve delivery speeds and reduce costs.

Healthcare companies analyze medical test data and reported symptoms to help doctors diagnose diseases earlier and treat them more effectively.

How Data Science Is Conducted

The process of analyzing and acting upon data is iterative rather

than linear, but this is how the work typically flows for a data modeling project:



Machine Learning

7.1 What Is Machine Learning?

In the past 30 years there has been an explosion of data. This mass of data is useless unless we analyze it and find the patterns hidden within the data. Machine Learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

7.2 Two Approaches To Machine Learning

7.3 Supervised Machine Learning

Supervised machine learning algorithms are the most commonly used. With this model, a data scientist acts as a guide and teaches the algorithm what

conclusions it should make. Just as a child learns to identify fruits by memorizing them in a picture book, in supervised learning, the algorithm is trained by a dataset that is already labeled and has a predefined output. Examples of supervised machine learning include algorithms such as linear and logistic regression, multiclass classification, and support vector machines.

7.4 Unsupervised Machine Learning

Unsupervised machine learning uses a more independent approach, in which a computer learns to identify complex processes and patterns without a human providing close, constant guidance. Unsupervised machine learning involves training based on data that does not have labels or a specific, defined output. To continue the childhood teaching analogy, unsupervised machine learning is akin to a child learning to identify fruit by observing colors and patterns, rather thanmemgorizin the names with a teacher's help.

.

Recruiting Tools

8.1 Jupiter Notebook: The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

- **Advantages:**

I. Interactive Computing

II. Easy Data Visualization

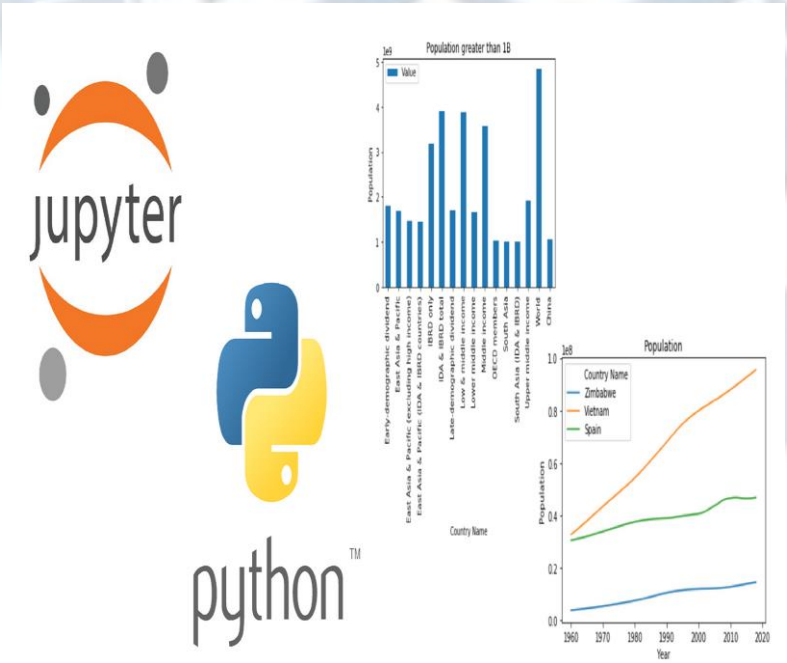
III. Reproductivity

IV. EDA

V. Library and Ecosystem

VI. Cloud Integration

Example-

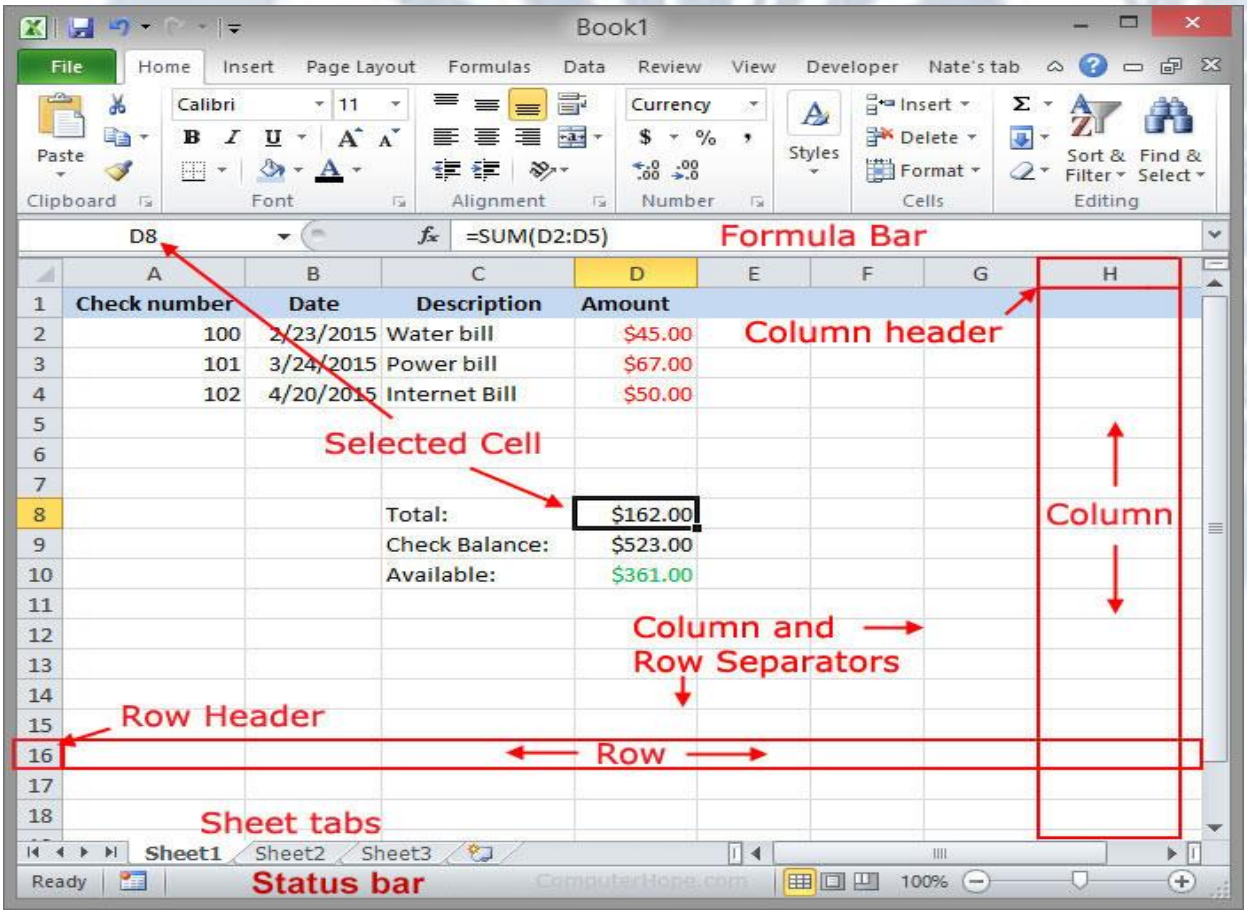


8.2 Excel: Microsoft Excel is one of the most popular applications for data analysis. Equipped with built-in pivot tables, they are without a doubt the most sought-after analytic tool available. It is an all-in-one data management software that allows you to easily import, explore, clean, analyze, and visualize your data. In this article, we will discuss the various methods of data analysis in Excel.

Advantages of Using Excel in Data Science

There are several advantages to using Excel in data science. Firstly, it is a widely available and affordable tool that can be used by businesses of all sizes. Secondly, it is a user-friendly tool that requires minimal training to use. Thirdly, Excel provides a wide range of features that can be used for data manipulation, analysis, and visualization. Finally, Excel integrates seamlessly with other tools and software, making it easy to incorporate into existing workflows.

Example-



Related Project During Internship

PTS PROJECT ON NETFLIX DATA ANALYSIS

BY- SHYAM BIHARI KUMAR

(SID-99695)

Objectives - Exploratory Data Analysis On netflix dataset

9.1 **Required Library:**

Numpy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

Import Method – import numpy as np

Pandas: What is Pandas? Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Import Method – import Pandas as pd

Seaboen: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Import Method - import seaborn as sns

Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.

Import Method - import matplotlib.pyplot as plt

Dataset_ Download:

https://github.com/Shyamsaurabh325/NETFLIXDATASETEDA/blob/main/netflix_titles.csv

About Dataset: The raw data is Web Scrapped through Selenium. It contains Unlabelled text data of around 9000 Netflix Shows and Movies along with Full details like Cast, Release Year, Rating, Description, etc.

Objectives - Exploratory Data Analysis On netflix dataset

In [30]:
```
# Now Inport the required Library
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

Data Loading

In [31]:
```
# Read the csv file
netflix=pd.read_csv("C:\\Users\\Shyam Bihari Kumar\\OneDrive\\Desktop\\All Data Set\\netflix_titles.csv")
```

Description:

First Import the all required library and after that load csv dataset file after that perform any task on this project

In [32]:
```
# Find out 5 rows and column
netflix.head()
```

Out[32]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV	In a city of coaching centers

Description: Find out the head of data that means top 5 rows and top 5column

That is help to perform next task easy

Data Exploration: Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

During exploration, raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to visually explore data sets, look for similarities, patterns and outliers and to identify the relationships between different variables This is also sometimes referred to as exploratory data analysis, which is a statistical technique employed to analyze raw data sets in search of their broad characteristics

Why is data exploration important?

Humans are visual learners, able to process visual data much more easily than numerical data. Consequently, it's challenging for data scientists to review thousands of rows of data points and infer meaning without assistance.

Data visualization tools and elements like colors, shapes, lines, graphs and angles aid in effective data exploration of metadata, enabling relationships or anomalies to be detected.

Data Exploration

In [43]:
```
# Finding the data information related to nul or object
netflix.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

In [44]:
```
# Find the shape of dataset (all rows and column)
netflix.shape
```

Out[44]:
```
(8807, 12)
```

In [45]:
```
# Find the assign value is null or not
print(netflix.duplicated().sum(),netflix.isnull().sum(),sep = "\n\n")
```

```
0

show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

Description: In this data frame cheque the Null_Value and data types available in this dataset after that we cheque shape of data

Data Cleaning: Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

How to clean data:

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find "N/A" and "Not Applicable" both appear, but they should be analyzed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember

Step 4: Handle missing data

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.

As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.

As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

Does the data make sense?

Does the data follow the appropriate rules for its field?

Does it prove or disprove your working theory, or bring any insight to light?

Can you find trends in the data to help you form your next theory?

If not, is that because of a data quality issue?

Advantages and benefits of data cleaning

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

Removal of errors when multiple sources of data are at play.

Fewer errors make for happier clients and less-frustrated employees.

Ability to map the different functions and what your data is intended to do.

Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.

Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

Data Clining

In [7]:
```
netflix["date_added"] = pd.to_datetime(netflix["date_added"])
netflix.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   show_id       8807 non-null   object        
 1   type          8807 non-null   object        
 2   title         8807 non-null   object        
 3   director      6173 non-null   object        
 4   cast          7982 non-null   object        
 5   country       7976 non-null   object        
 6   date_added    8797 non-null   datetime64[ns]
 7   release_year  8807 non-null   int64         
 8   rating        8803 non-null   object        
 9   duration      8804 non-null   object        
 10  listed_in     8807 non-null   object        
 11  description   8807 non-null   object        
dtypes: datetime64[ns](1), int64(1), object(10)
memory usage: 825.8+ KB
```

In [8]:
```
netflix["date_added"] = pd.to_datetime(netflix["date_added"])
netflix["month"] = netflix["date_added"].dt.month
netflix["year"] = netflix["date_added"].dt.year
netflix.head(3)
```

Out[8]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	month	year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	9.0	2021.0
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	9.0	2021.0
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	9.0	2021.0

In [9]:
```
netflix = netflix.rename(columns = {"listed_in":"Genre"})
netflix["Genre"] = netflix["Genre"].apply(lambda x : x.split(",")[0])
netflix.head(3)
```

Out[9]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	Genre	description	month	year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	9.0	2021.0
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows	After crossing paths at a party, a Cape Town t...	9.0	2021.0
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows	To protect his family from a powerful drug lor...	9.0	2021.0

In [10]:
```
netflix.fillna({"director":"Missing","cast":"Missing","country":"Unavailable","date_added":"Unavailable","rating":"Unavailab
                "duration":"Unavailable","month":"Unavailable","year":"Unavailable"},inplace= True)
netflix.isnull().sum()
```

Out[10]:
```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
Genre           0
description     0
month           0
year            0
dtype: int64
```

Description: In this data frame we see all null values is show o by performing data cleaning

Data Visualization: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Advantages of Visualization

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

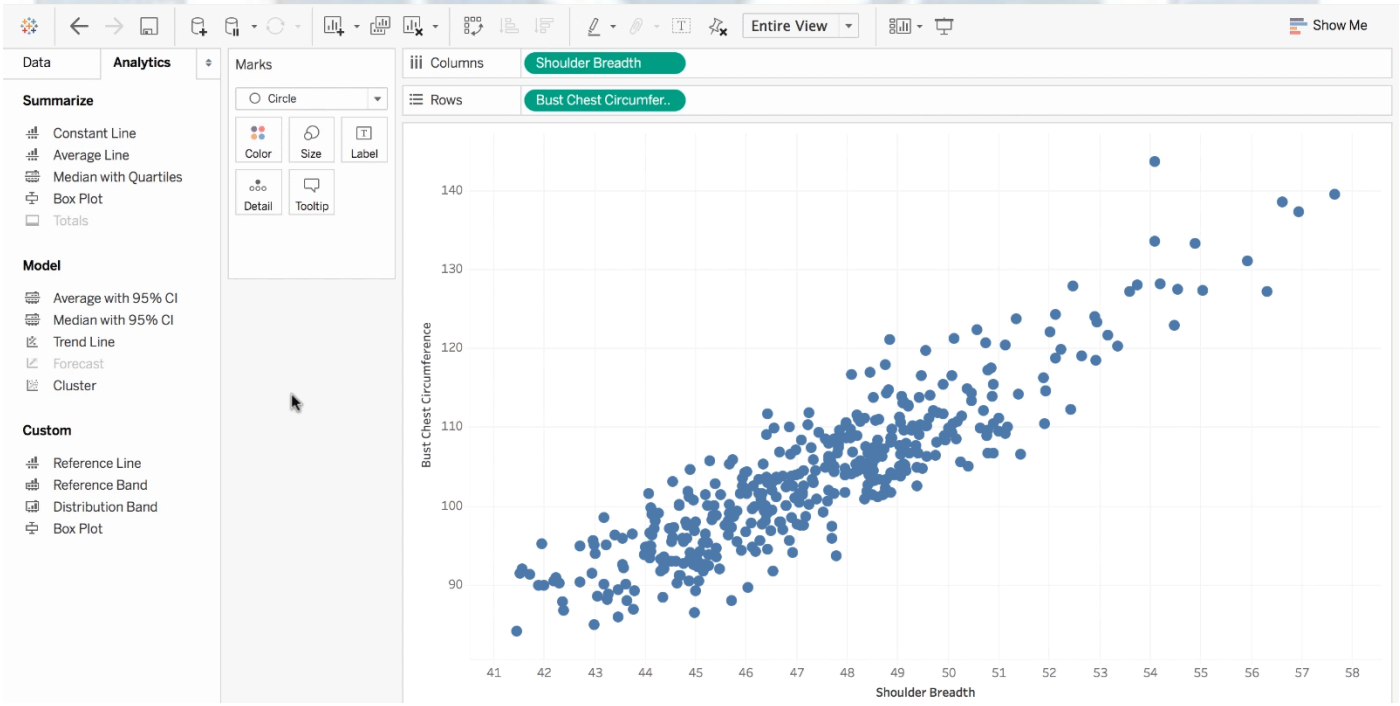
Some other advantages of data visualization include:

Easily sharing information.

Interactively explore opportunities.

Visualize patterns and relationships.

Example:



Visualization

How many Movies & TV Shows are in the dataset

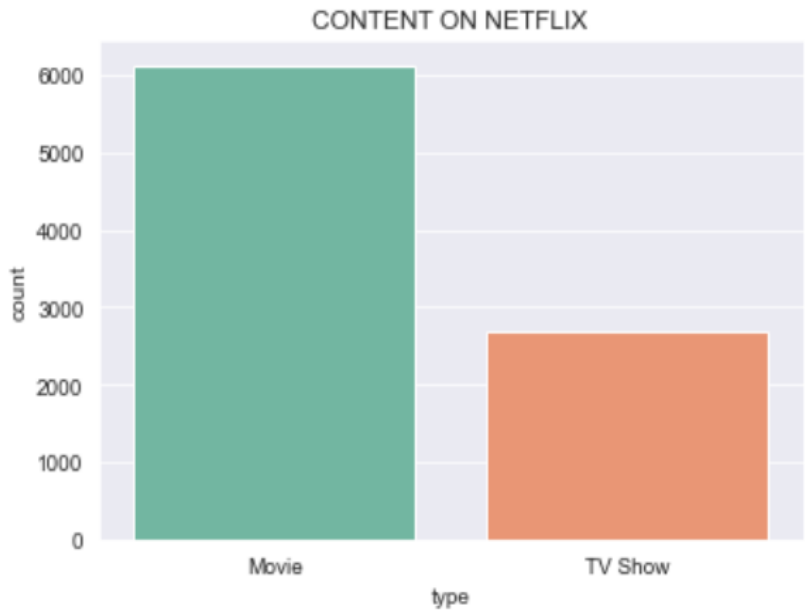
In [11]:
```
1 netflix["type"].value_counts()
```

Out[11]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

Count the value of total content Movie and TV show

In [12]:
```
1 plt.figure(figsize =(6,4))
2 sns.set_style("darkgrid")
3 sns.countplot(x = "type",data =netflix, palette = "Set2")
4 plt.title("CONTENT ON NETFLIX")
```

Out[12]:
```
Text(0.5, 1.0, 'CONTENT ON NETFLIX')
```



In which year there was highest no of Tv Shows and Movie

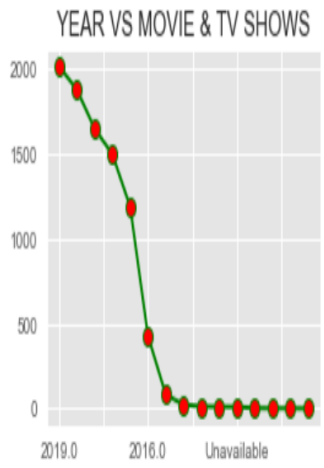
In [13]:
```
netflix["year"].value_counts()
```

Out[13]:
```
2019.0          2016
2020.0          1879
2018.0          1649
2021.0          1498
2017.0          1188
2016.0           429
2015.0            82
2014.0            24
2011.0            13
2013.0            11
Unavailable       10
2012.0             3
2009.0             2
2008.0             2
2010.0             1
Name: year, dtype: int64
```

FIND VALUES YEAR VS MOVIE & TV SHOWS

In [58]:
```
plt.figure(figsize =(4,3))
plt.style.use("ggplot")
netflix["year"].value_counts().plot(kind ="line",color = 'green',linestyle = 'solid', marker = 'o',markerfacecolor = 'red',
plt.title("YEAR VS MOVIE & TV SHOWS")
```

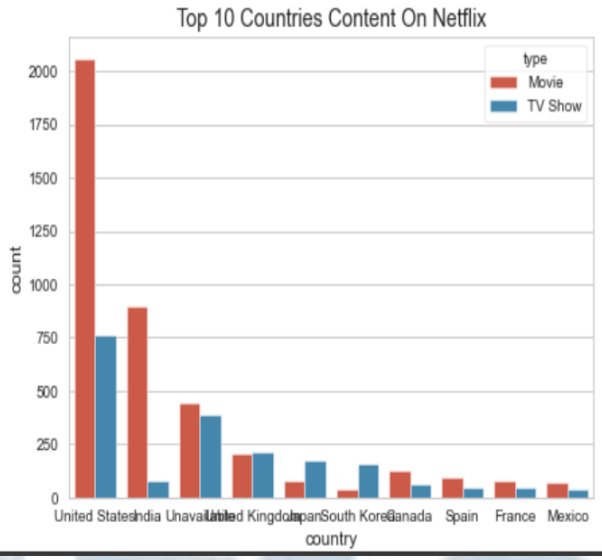
Out[58]: Text(0.5, 1.0, 'YEAR VS MOVIE & TV SHOWS')



Top 10 Countries Content On Netflix

In [60]:
```
plt.figure(figsize = (7,5))
sns.set_style("whitegrid")
sns.countplot(x="country", order = netflix["country"].value_counts().index[0:10],hue = "type",data = netflix)
plt.title("Top 10 Countries Content On Netflix")
```

Out[60]: Text(0.5, 1.0, 'Top 10 Countries Content On Netflix')



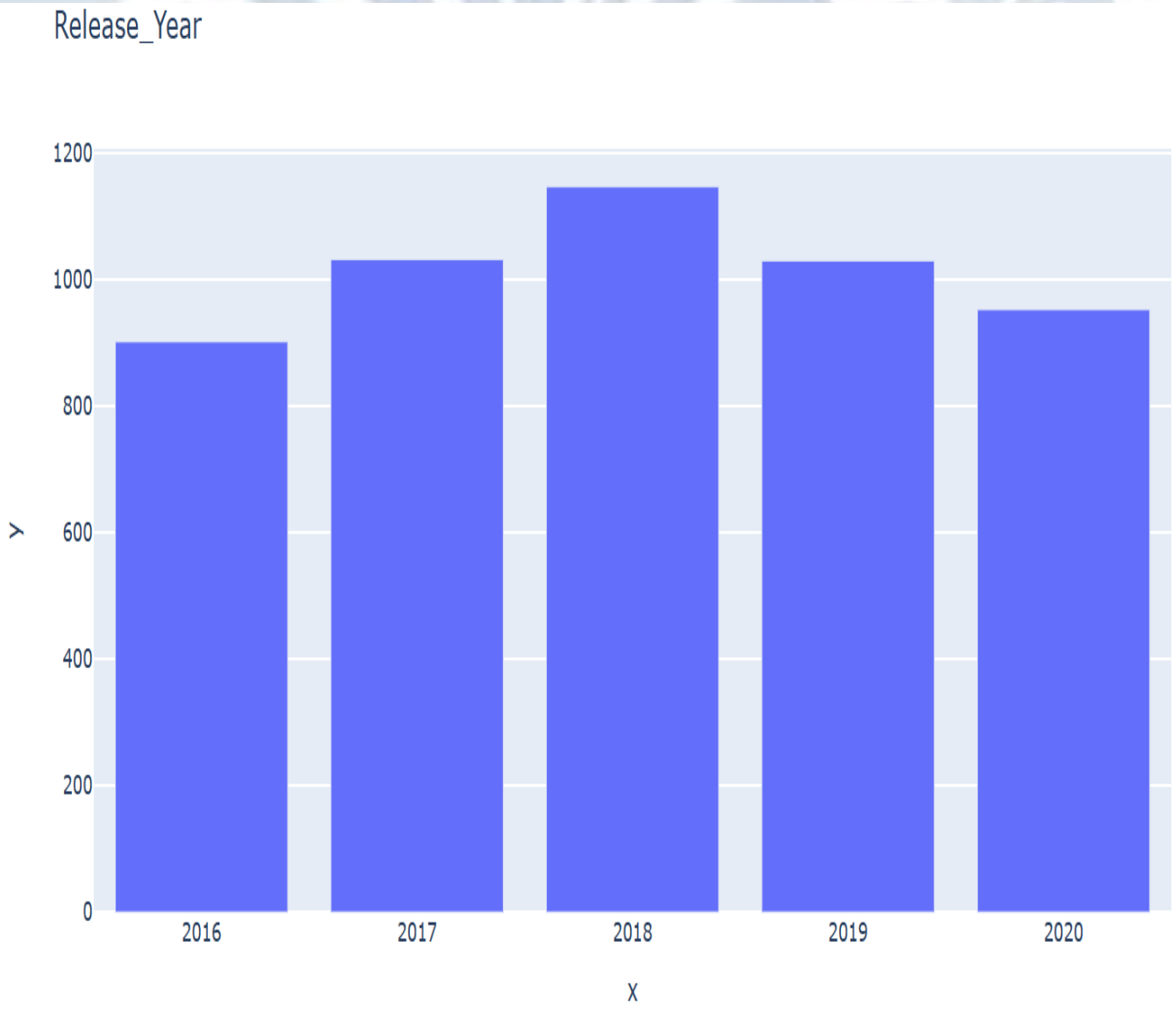
Now find the number of movie release in year by year

In [61]:
```
# Now find the number of movie release in year by year
r=netflix['release_year'].value_counts().nlargest(5)
r
```

Out[61]:
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
Name: release_year, dtype: int64
```

In [17]:

```
## fig.update_layout(xaxis=r.index,y=r)
fig=px.bar(x=r.index,y=r,title='Release_Year')
fig.show()
```



Now find the number of movie release by country

In [63]:
```
# Now find the number of movie release by country
c=netflix['country'].value_counts().nlargest(5)
c
```

Out[63]:
```
United States     2818
India              972
Unavailable        831
United Kingdom     419
Japan              245
Name: country, dtype: int64
```

In [64]:
```
c.index
```

Out[64]:
```
Index(['United States', 'India', 'Unavailable', 'United Kingdom', 'Japan'], dtype='object')
```

In [65]:
```
c.values
```

Out[65]:
```
array([2818,  972,  831,  419,  245], dtype=int64)
```

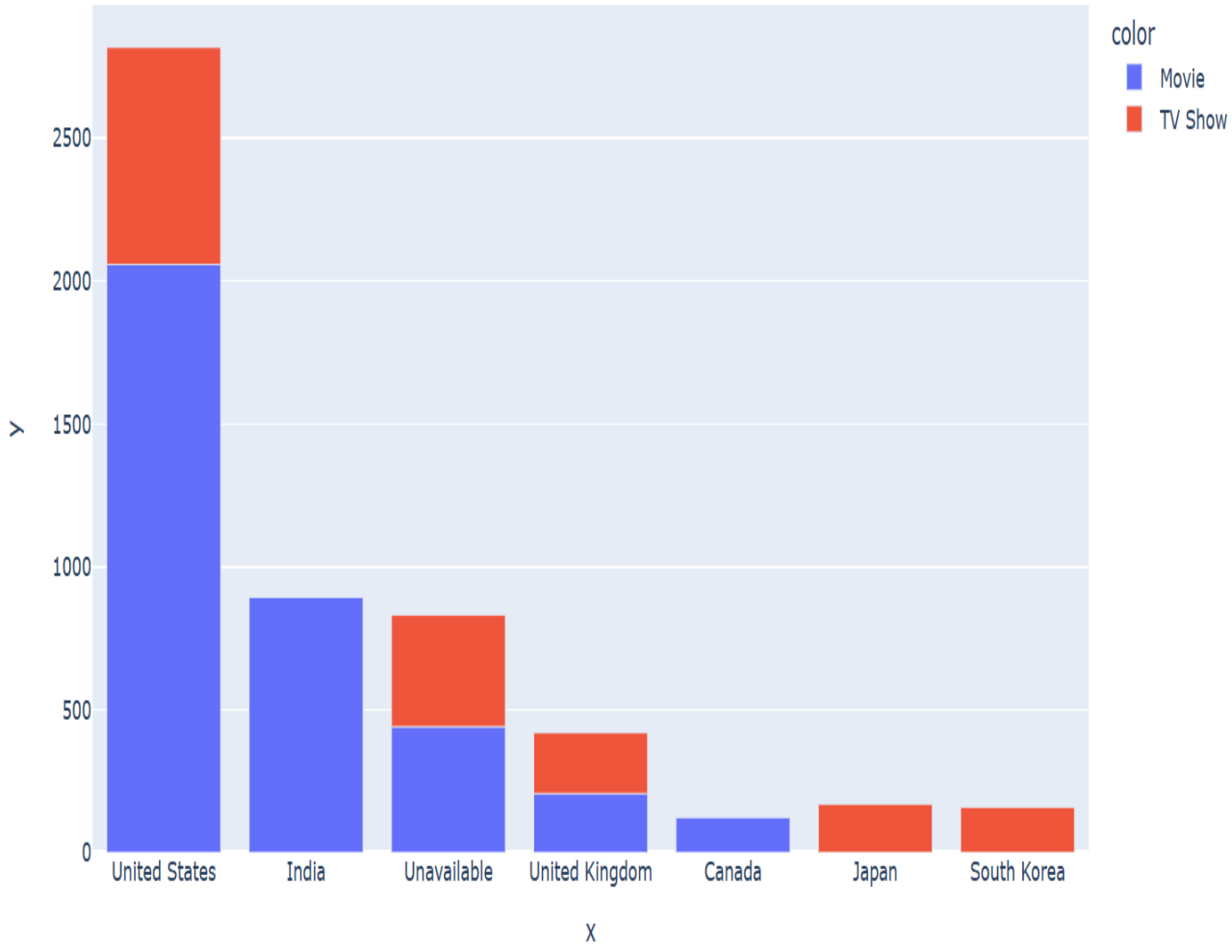
In [66]:
```
n=netflix.groupby('country')['type'].value_counts().nlargest(10)
n1=n.index.map(lambda x:x[0])
n2=n.index.map(lambda x:x[1])
n
```

Out[66]:
```
country         type
United States   Movie      2058
India           Movie       893
United States   TV Show     760
Unavailable     Movie       440
                TV Show     391
United Kingdom  TV Show     213
                Movie       206
Japan           TV Show     169
South Korea     TV Show     158
Canada          Movie       122
Name: type, dtype: int64
```

The above plot shows top 10 countries with maximum no. of movies i.e 0 and tv shows i.e

2]:

```
1 fig=px.bar(x=n1,y=n,color=n2)
2 fig.show()
```



The above dataframe gives information of movies by David Fincher

In [24]:
```
1 netflix[(netflix['director']=='David Fincher')]
```

Out[24]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	Genre	description	month	year
600	s601	Movie	The Game	David Fincher	Michael Douglas, Sean Penn, Deborah Kara Unger...	United States	2021-07-01 00:00:00	1997	R	129 min	Thrillers	An aloof investment banker's life spirals into...	7.0	2021.0
1595	s1596	Movie	MANK	David Fincher	Gary Oldman, Amanda Seyfried, Charles Dance, L...	United States	2020-12-04 00:00:00	2020	R	133 min	Dramas	1930s Hollywood is reevaluated through the eye...	12.0	2020.0
7701	s7702	Movie	Panic Room	David Fincher	Jodie Foster, Forest Whitaker, Dwight Yoakam, ...	United States	2019-08-01 00:00:00	2002	R	112 min	Thrillers	A woman and her daughter are caught in a game ...	8.0	2019.0
8320	s8321	Movie	The Girl with the Dragon Tattoo	David Fincher	Daniel Craig, Rooney Mara, Christopher Plummer...	United States, Sweden, Norway	2021-01-05 00:00:00	2011	R	158 min	Dramas	When a young computer hacker is tasked with in...	1.0	2021.0
8511	s8512	Movie	The Social Network	David Fincher	Jesse Eisenberg, Andrew Garfield, Justin Timbe...	United States	2020-04-01 00:00:00	2010	PG-13	121 min	Dramas	Director David Fincher's biographical drama ch...	4.0	2020.0

The above dataframe shows the information of movies originated from united states

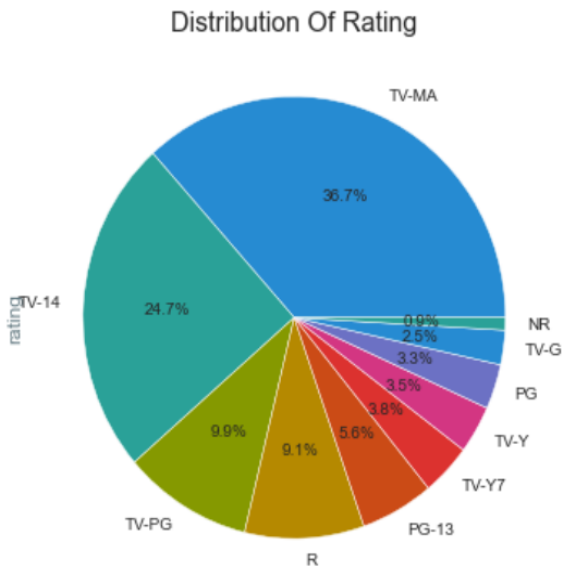
In [25]:
```
1 netflix[(netflix['country']=='United States')]
```

Out[25]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	Genre	description	month	year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Missing	United States	2021-09-25 00:00:00	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	9.0	2021.0
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	2021-09-24 00:00:00	2021	PG-13	104 min	Comedies	A woman adjusting to life after a loss contend...	9.0	2021.0
15	s16	TV Show	Dear White People	Missing	Logan Browning, Brandon P. Bell, DeRon Horton,...	United States	2021-09-22 00:00:00	2021	TV-MA	4 Seasons	TV Comedies	Students of color navigate the daily slights a...	9.0	2021.0
27	s28	Movie	Grown Ups	Dennis Dugan	Adam Sandler, Kevin James, Chris Rock, David S...	United States	2021-09-20 00:00:00	2010	PG-13	103 min	Comedies	Mourning the loss of their beloved junior high...	9.0	2021.0
					Keri Russell,							A family's		

In [26]:
```
plt.figure(figsize = (10,6))
plt.style.use("Solarize_Light2")
netflix["rating"].value_counts()[:10].plot(kind ="pie",autopct= "%1.1f%%")
plt.title("Distribution Of Rating")
```

Out[26]: Text(0.5, 1.0, 'Distribution Of Rating')



The above dataframe shows information of movies with rating PG-13

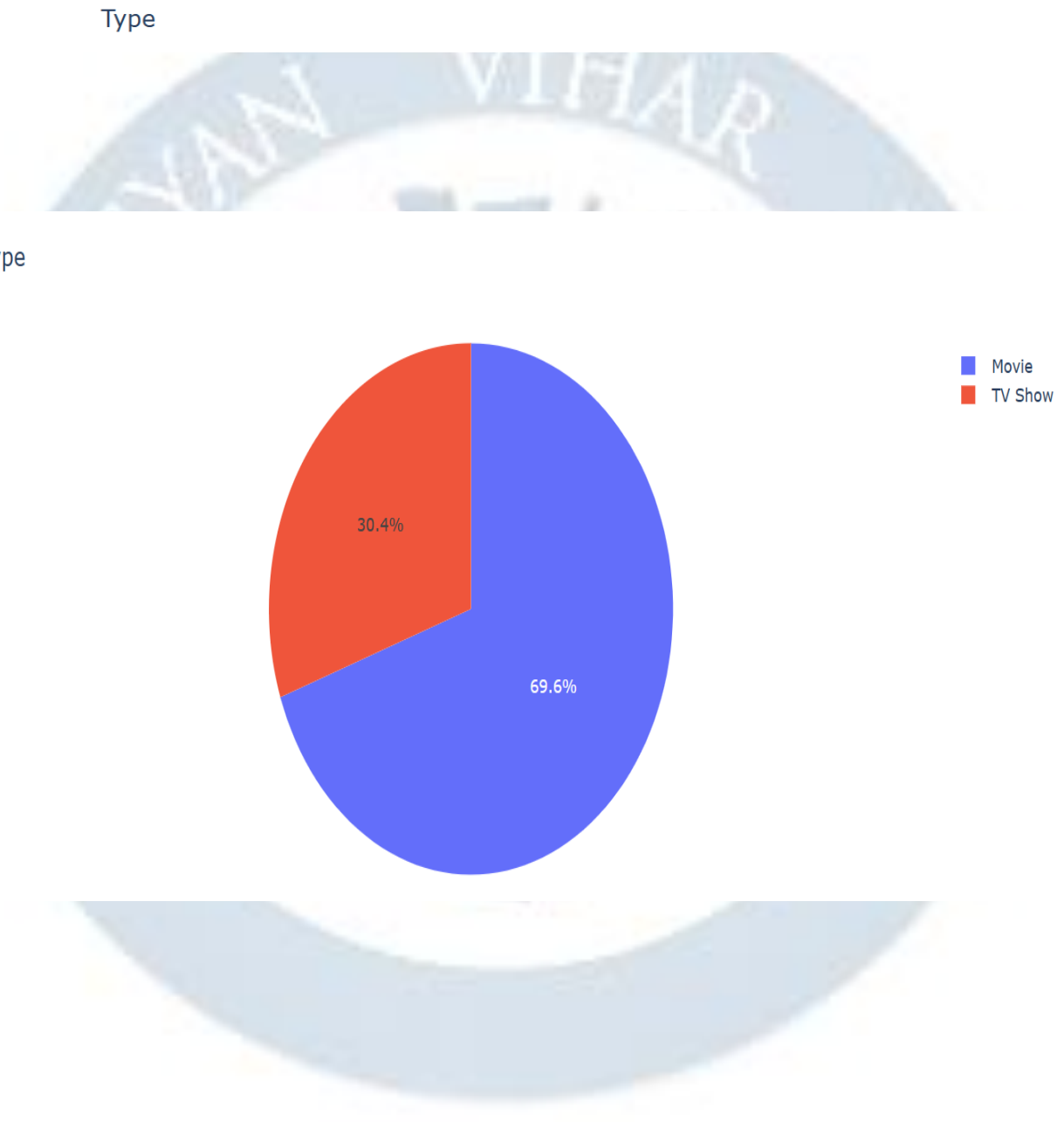
]:
```
netflix[(netflix['rating']=='PG-13')]
```

]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	Genre	description	month	year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Missing	United States	2021-09-25 00:00:00	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	9.0	2021.0
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	2021-09-24 00:00:00	2021	PG-13	104 min	Comedies	A woman adjusting to life after a loss contend...	9.0	2021.0
27	s28	Movie	Grown Ups	Dennis Dugan	Adam Sandler, Kevin James, Chris Rock, David S...	United States	2021-09-20 00:00:00	2010	PG-13	103 min	Comedies	Mourning the loss of their beloved junior high...	9.0	2021.0
28	s29	Movie	Dark Skies	Scott Stewart	Keri Russell, Josh Hamilton, J.K. Simmons, Dak...	United States	2021-09-19 00:00:00	2013	PG-13	97 min	Horror Movies	A family's idyllic suburban life shatters when...	9.0	2021.0
29	s30	Movie	Paranoia	Robert Luketic	Liam Hemsworth, Gary Oldman, Amber Heard,	United States, India,	2021-09-19 00:00:00	2013	PG-13	106 min	Thrillers	Blackmailed by his company's CEO, a low-	9.0	2021.0

Now find the percentage of releasing movie and tv shows

In [23]:
```
# The above plot shows 69.6% of the data is about movies and 30.4% is about tv shows
fig=px.pie(netflix,names='type',title='Type')
fig.show()
```



Conclusion

The dataset has 8807 rows and 11 columns..

Out of 8807 entries 6131 are TV shows.

2018 have the maximum releases of movies and Tv shows.

united states has released maximum no. of movies and tv shows on netflix.

Top 10 countries with maximum no. of movies and tv shows.

information of movies directed by David Fincher.

the information of movies originated from united states.

information about all the comedy movies and tv shows.

information of movies with rating PG-13.

information of movies in which Damon Wayans has acted.

Dramas and international movies has more records in dataset.

Most of the movies and Tv shows from dataset are rated TV-MA.

Rajiv chilaka has directed maximum movies and tv shows.

69.6% of the data is about movies and 30.4% is about tv shows.

Future Scope

The future scope of a Netflix data analysis project can be quite extensive, given the ever-evolving nature of the entertainment industry and the wealth of data that Netflix generates. Here are some potential areas for future development and expansion:

1.Content Recommendation and Personalization:

Enhancing the recommendation algorithms to provide more personalized and relevant content recommendations to users.

Incorporating real-time user feedback and interactions to improve recommendation accuracy.

Exploring new data sources, such as social media activity, to further refine content recommendations.

2.Content Creation and Acquisition:

Analyzing viewer data to inform content creation and acquisition decisions. This could involve predicting which types of content will be successful and tailoring productions accordingly.

Identifying trends and gaps in the content library to guide future content investments.

3.Content Quality Assessment:

Developing algorithms to assess and improve the quality of content by analyzing viewer ratings, reviews, and engagement metrics.

Utilizing sentiment analysis and natural language processing to gauge audience reactions and feedback.

4.User Engagement and Retention:

Focusing on strategies to increase user engagement, reduce churn, and improve the overall user experience.

Analyzing user behavior and feedback to identify pain points and opportunities for improvement.

5.Content Localization and Global Expansion:

Using data analysis to tailor content for specific markets and cultures, optimizing localization efforts.

Identifying new markets for expansion based on user demographics and interests.

Pricing and Monetization Strategies:

Analyzing pricing data and user behavior to optimize subscription plans and pricing strategies.

Experimenting with new monetization models, such as tiered pricing or bundled services.

6.User Profiling and Segmentation:

Developing more sophisticated user profiles and segmentation techniques to target specific user groups with tailored content and marketing.

Utilizing machine learning and clustering algorithms for more accurate user grouping.

7.Content Delivery and Infrastructure Optimization:

Improving the efficiency of content delivery and streaming to enhance the overall viewing experience.

Analyzing network and infrastructure data to reduce latency and optimize streaming quality.

8.Security and Anti-Piracy Measures:

Utilizing data analysis to identify and combat piracy and unauthorized sharing of content.

Implementing security measures to protect user data and privacy.

9.Regulatory Compliance:

Ensuring compliance with evolving data privacy regulations and content rating guidelines.

Adapting data management practices to meet changing legal requirements.

10.Research and Development:

Investing in research to explore emerging technologies, such as virtual reality (VR) and augmented reality (AR) for content delivery.

Experimenting with new data sources and analysis techniques to stay at the forefront of the industry.

Reference

If you're looking for authoritative references in this field, here are some well-regarded books and online resources that cover data science and machine learning with Python:

Books:

"Python for Data Analysis" by Wes McKinney: This book focuses on practical data analysis using Python and the Pandas library.

"Introduction to Machine Learning with Python" by Andreas C. Müller & Sarah Guido: An excellent resource for those new to machine learning.

"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron: A practical guide to machine learning using Python libraries.

"Python Machine Learning" by Sebastian Raschka & Vahid Mirjalili: Covers a wide range of machine learning topics with practical examples.

"Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville: A comprehensive resource for deep learning and neural networks.

Online Resources:

Coursera and edX: These platforms offer courses on data science and machine learning, often using Python. You can find courses from top universities and institutions.

Kaggle: Kaggle provides datasets, competitions, and tutorials for data science and machine learning. It's an excellent platform to practice and learn.

Towards Data Science: A Medium publication with many articles and tutorials on data science and machine learning using Python.

DataCamp: Offers a variety of online courses in data science and machine learning using Python.

scikit-learn Documentation: The official documentation for the scikit-learn library provides detailed information on using scikit-learn for machine learning.

Please note that the field of data science and machine learning is rapidly evolving, so it's essential to look for the most up-to-date resources. Additionally, your specific interests and skill level will determine which resources are most suitable for your learning journey.