Netflix Exploratory Data Analysis (Netflix EDA)
Project Report

Group Number - 1

Name	Reg. No.
Aakash Chaurasiya	20BCS001
Abhishek Dubey	20BCS002
Aditya Tiwari	20BCS007
Devansh Mahant	20BCS039
Tushar Dhotre	20BCS042
Kartik Bhamare	20BCS066
Pruthviraj digambar kamble	20BCS104
Mahendra singh puniya	20BCS082

Under the Guidance of,
Uma mam
Department of Computer Science,
Indian Institute of Information Technology,
Dharwad

Abstract

In this report our focus is mainly on analysing the data and various factors affecting the data available on Netflix. Data visualization was a primary aim and was implemented using pandas, matplotlib, bar_chart_race, seaborn, wordcloud and stopwords. All these python graphing libraries make interactive, publication-quality graphs. Our project does an exploratory data analysis of the Netflix dataset.

The term “Exploratory and Sentiment Analysis” is a conjunction of two separately unique approaches present in the vast field of Data Science. The key to this project is to enhance the value of the data being utilized, in our case it is Netflix data – which is an open-source data set obtained from Kaggle – that was wrangled to derive maximum insights using exploratory data analysis. This project introduces systematic and insightful usage of various techniques for exploratory data analysis by utilizing various packages.

Contents

Introduction

Exploratory Data Analysis or EDA as it is commonly called is a process or stage in any data science project that cannot be overlooked or talked about enough.

This is where the data scientist “gets a feel or understands” the data he/she wants to build a model on.

In cases where the end product of such a project isn’t some ML or AI product, EDA can result in great insights and recommendations about business problems through pattern discovery, hypothesis testing, and checking of assumptions. All these are usually achieved with help of *summary statistics and data visualizations*.

The ultimate aim here is to aid effective and efficient decision making which may affect businesses positively.

In this article, we are not going to be deploying any model to create an ML or AI product after our EDA. We’re simply going to use EDA to “explore” and gain an understanding of a dataset containing Netflix’s contents between 2008 and 2020. We’re basically going to be using EDA to get a “quick summary” of what the dataset contains and to gain insights on the trends in viewing of movies and T.V. shows on the streaming platform.

Report on Netflix EDA (Present Investigation)

EDA as a method of investigating Data

EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Software Packages used for EDA

A Python package usually consists of several modules and modules are a file containing Python definitions and statements. A module can define functions, classes, and variables.

So the question arises, why do we use packages or modules for exploratory data analysis?
The answer to this is that packages and modules allow us to logically organize our Python code. Grouping related code into a module makes the code easier to understand and use.

In our project we have used several packages and modules which help us in handling data and creating visualisations eg. pandas, matplotlib, bar_chart_race, seaborn, wordcloud, STOPWORDS etc.

Pandas provides many useful functions to inspect only the data we need. Pandas is mainly used for data processing, data slicing and data cleaning. We can use functions like df.head(n) to get the first n rows or df.tail(n) to print the last n rows, df.iloc[x,y] to display particular columns or rows and many more functions which are a part of Pandas, as required by us.

We used matplotlib package which is a comprehensive library for creating static, animated, and interactive visualizations from data.

We used bar_chart_race packages to get an animated bar chart and used wordcloud and STOPWORDS to create a word cloud in which the size of each word indicates its frequency or importance.

Data Cleaning

Data cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set's accuracy without necessarily deleting information.

When it comes to data structure, it is not improbable that data may contain incomplete, inconsistent or missing values. If the data is irrelevant or error-prone then it leads to an incorrect model building which may hinder the process or provide inaccurate results.

For example, listing movies and TV shows on Netflix involves the study of title, cast, release date and many more, so basically it may contain some null values too. Now you want to know which country produces the least movies or which director produces more romantic TV shows. But if the data is corrupted or contains missing values then the insights derived from the data might be incorrect leading one to be misguided in the decision making process and which might land one in trouble.

Data Summarisation and Summary Statistics

The term Data Summarization refers to presenting the summary of generated data in an easily comprehensible and informative manner.

Presenting the raw data (the data that was generated which is essentially the entire dataset of individual measurements) is not practical in many cases.

For example, the listing of movies and tv shows on Netflix involves the study of title, cast, release date and many more. Presenting such complex data would need several printed pages, and convey no easily comprehensible information.

A carefully chosen summary of raw data would convey many trends and patterns of the data in an easily accessible manner. The term ‘data mining’ refers to exactly this; extracting meaningful information from the raw data.

For example, what are the top genres or top countries with the most movies and TV shows? The way data is presented is a very important, although often overlooked aspect in statistics. Data summarization comes much before any statistical tests; indeed choosing appropriate statistical tests depends on the general trends of the data revealed in the summarization step.

We have used many functions, methods and attributes to summarize our data set like:- df.type(), df.shape, df.columns etc.

Data Visualization

Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

Visualization helps you monitor events or activities at a glance by providing insights on one or more pages or screens. An interactive visualization makes it easy to sort, filter, or drill into different types of data as needed. Visualization techniques can be used to identify what is happening, why it's happening, and what will happen next at speed.

The more creative we become with data, the more insights we can visualize.

For example, we have a listing of movies and TV shows on Netflix involving the study of title, cast, release date and many more. So when we require insights like what are the top genres or top countries with the most number of movies and TV shows, the way data is presented is very important and visualisation is the best way.

We have used various types of visualization techniques to gain insights from our dataset like bar chart, pie chart, scatter plot, etc.

Summarization techniques and Summary Statistics used for investigating the dataset

1. type()

   The type() function returns the type of the object which is passed as a parameter to the function.

2. df.shape

   The shape attribute returns a tuple of the shape of the underlying data for the dataframe object, i.e., the number of rows and columns respectively of the dataframe.

3. df.columns

   The columns attribute returns the column labels of all the columns in the dataframe.

4. df.head()

   The head() method returns the top n (5 if no parameter is passed) rows of the dataframe, where n is the parameter passed to the method.

5. df.tail()

   The tail method returns the bottom n (5 if no parameter is passed) rows of the dataframe, where n is the parameter passed to the method.

6. df.info()

   The info() function is used to get a concise summary of the dataframe. It returns information like the number of non null values and the data type for each column in the dataframe.

7. df.describe()

   The describe() method is used for calculating statistical data like percentile, mean and std of the numerical values of the numeric columns of the dataframe.

Data cleaning techniques used for investigating the dataset

1. We have used many techniques for data cleaning in our project like date is in the form of string and thus many times it is important to convert the date to proper format, i.e. datetime object. For this we can use :-

```
df['date_added']= pd.to_datetime(df['date_added'].str.strip(),
format= "%B %d, %Y")
```

The above code cell changes the dates in the date_added column to the datetime format, with the rows ordered from latest to earliest date in the date_added column.

2. Most developers want to know about null values or missing data in their datasets. Within pandas, a missing value is denoted by NaN. In order to check null values in Pandas Dataframe we can use :-

```
df.isnull().sum().sort_values(ascending=False)
```

The above code cell prints the total number of null values which are present in all the columns of the data frame and presents them in descending order.
From this information we understand that the director, country and cast columns have the most number of null values/missing data.

3. If we want to replace NaN value with any other value we can use pandas series.fillna(). This is used to fill NaN values using the value passed as a parameter in the fillna function. We used this technique on the results dataframe, in the code for the bar chart animation as follows :-

```
result=result.fillna(0)
```

The above code cell replaced all NaN values with 0.

4. The pandas duplicated() method helps in analyzing duplicate values only. It returns a boolean series which is False only for all unique elements. If we want to know any duplicate values in a particular column we can use :-

```
df['show_id'].duplicated().any()
```

The result of the above code cell tells us that none of the show ids have been repeated in the dataset and therefore no movie or TV show has been duplicated.

Data visualisation techniques used for investigating the dataset

1. Vertical Bar Graph

   A vertical bar graph is the most common type of bar chart and it is also referred to as a column graph. It represents the numerical value of research variables using vertical bars whose lengths are proportional to the quantities that they represent.

2. Horizontal Bar Graph

   A Horizontal bar graph uses horizontal bars to represent the numerical values of research variables, whose lengths are proportional to the quantities that they represent. They are often used when there are a large number of research variables or when the lengths of the labels of the research variables are long and cannot fit in a vertical bar graph.

3. Grouped Bar Graph

   A grouped bar graph is used when each of the research variables has two or more sub variables which need to be measured separately with a bar plot of their own. They can be either horizontal or vertical.

4. Animated Horizontal Bar Graph

   The animated bar chart helps us visualize the change in trends over time. These types of charts are very popular as they provide a holistic data story/insight in a concise and easy-to-understand chart.

5. Pie Chart

   Pie charts show the research variables as part of a whole, i.e., they show the proportion of each variable. Pie charts are circular charts with each sector of the pie chart representing a variable.

6. Scatter Plot

   A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. It uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

7. Line Graph

   A line graph is a type of chart used to show information that changes over time. We plot line graphs using several points connected by straight lines. Time varies along the x-axis while the other variables vary along the y-axis.

8. Word Cloud

   Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

Results and Discussion

1. The Netflix streaming platform has more than 6000 movies and a little less than 3000 T.V. shows.

2. More than 2500 movies and tv shows originated from the United states, nearly 1000 from india,less than 500 from uk japan and south korea.

3. A large count of Netflix content is made with a “TV-14” rating. But the largest count of T.V. shows have a “TV-MA” rating. The 3rd and 4th largest count of Netflix content is made with “TV-PG” and “R” rating respectively. And only a few movies are rated as NC-17.

4. The Netflix streaming platform contains more than 2000 movies and more than 1000 T.V. shows with TV-MA rating, little less than 1500 movies and nearly 750 T.V. shows are rated as TV-14, very few movies and T.V. shows are rated as NC-17.

5. Netflix has a few movies from the year 1940 to 1950 with short durations but from 1950 to 2000 more movies with increased durations were produced and from 2000 to 2020 the number of movies have significantly increased having more number of movies inclined towards smaller duration.

6. The number of TV-MA movies has increased significantly with time and that the range of the duration of TV-MA movies is bordered that of PG-13 movies, that is they can be very short or very long.while PG-13 movies tend to have a medium duration.

7. The most popular director on Netflix, with the most titles, is mainly international.

8. The highest number of tv shows and movies were released from 2016 to 2020. With more than 1000 TV shows and movies released in 2018 alone. With little less than 1000 movies and TV shows in 2017 and 2019.

9. Anupam kher has appeared in more than 40 movies that were released on Netflix. Followed by Shah Rukh Khan and Naseeruddin Shah with nearly 35 movies.
   And Boman Irani and Rupa Bhimani had the least movie apperence with little over 25 movies.

10. Rajiv Chilaka and Jan Suter directed more than 20 movies/T.V. shows released on Netflix, followed by Raul Campos directing less than 20 movies. At the 10th position we have Jay Chapman with around 12 movies/T.V. shows

11. Around 25 number of Netflix T.V. shows were done by Takahiro Sakurai, Yuki Kaji standing in second place with nearly 20 T.V. shows with Hiroshi at the 10th position with more than 10 shows.

12. Crime genre and kids T.V. shows were the most popular genres with nearly 400 movies and TV shows on netflix few of the least popular genres were spanish language T.V. shows, LGBTQ Movoes and sports movies and TV shows

13. There's been a spike in the number of TV shows and movies from the year 2000 to 2020 with the number of TV shows reaching upto 400 and number of movies little less than 800. With total content going as high as 1200.

Summary and Conclusion

"Listing Of Movies and TV Shows on Netflix " data set contains the listing of both movies and TV shows and various information about both like their titles with director , cast , country produced , release date on netflix , genre type , rating and many more .

It's a huge data set with around 8k+ rows and by just glancing that we can't get any useful information from it. That's why we have to perform operations . and in some columns containing no values we have to filter them too .

we have done various operations on our data to gain as many as insight information we can like :-

1. data cleaning to identify the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying .

2. We have used many data summarisation operations to get or print information about a particular column or row , to get data type and many more .

3. We have used various visualizations to translate our large dataset to help you monitor events or activities at a glance by providing insights on one or more pages or screens.

As for by performing above operations we have drawn many interesting inferences from the dataset Netflix titles; here's a summary of the few of them:

a. The most content type on Netflix is movies,
b. The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly,
c. The country by the amount of the produces content is the United States,
d. The most popular director on Netflix , with the most titles, is Jan Suter.
e. International Movies is a genre that is mostly in Netflix.
f. The largest count of Netflix content is made with a "TV-MA" rating .

References

a. **Github links**

1. **MadhumithaKannan/EDA-of-Netflix-data-using-Python**
   https://github.com/MadhumithaKannan/EDA-of-Netflix-data-using-Python/blob/master/.ipynb_checkpoints/final-checkpoint.ipynb

2. **netflix-eda**
   https://github.com/dwiknrd/medium-code/tree/master/netflix-eda.

3. **Netflix - EDA**
   https://github.com/siddharth271101/Netflix/blob/master/Netflix%20-%20EDA.ipynb

b. Analytics Vidhya

1. **Netflix Movies and TV Shows — Exploratory Data Analysis (EDA) and Visualization Using Python.**

https://medium.com/analytics-vidhya/netflix-movies-and-tvshows-exploratory-data-analysis-eda-and-visualization-using-python-80753fcfcf7

c. Seaborn.heatmap

https://seaborn.pydata.org/generated/seaborn.heatmap.html.

e.Jovian

https://jovian.ai/swapnilg4u/netflix-analysis-course-project