EXPERIENCE PROJECT
ADVANCED DATA SCIENCE TOOLS AND TECHNIQUES

Daniela Cristina Bula Gonçalves
Master in Data Science
Anhembi Morumbi EAD Laureate

Sao Paulo, SP
2020

The experiment project was carried out with the data from the Netflix Movie and TV show.

The overall context of the study is Netflix's recommendation of Movies and TV shows of the year 1925 through 2020. The main objective of the study is the elaborated achievements.

The area of Data Science has been occupying more and more space in the teams and decision of large corporations. Just as data scientists must be able to generate insights based on a data set, often unstructured, going through some steps that are fundamental during the construction of a Data Science project.

In this way the project was prepared for this presentation with the necessary information to understand the development of this project.

According to Netflix, which started its business as a mail-order DVD rental service in the subscription model, it now has nearly 193 million subscribers worldwide. The focus of the company, since 2007, has been its service of streaming movies, docs and series. Netflix has always been a data driven company and is responsible for developing one of the most robust recommendation systems for its subscriber base. Your recommendation system uses machine learning and artificial intelligence to assess the history of what you've seen on the platform and make recommendations for new movies and series. The recommendation algorithms are the basis of Netflix, being one of the main attributes of the company's revenue model, and they begin to act from the first access to the platform. Several data are taken into account by the algorithms, among them are:

Interactions with the platform: history of what was assisted and evaluations;
profile other users with similar preferences;
Information about the titles offered: genre, categories, actors, year of release, directors, Reviews.
User behavior: day of the week and time of day when they access, devices used (such as computer, tablet, smartphone), time on the platform, language, geographic region, etc.

The techniques used in Phase 1 and 2

Phase 1:

As we have seen, Netflix is known for its strong recommendation mechanisms. They use machine learning and a mix of collaborative, content-based filtering models to recommend TV shows and movies. At this stage we analyzed a data set of TV shows and movies available on Netflix to identify patterns from exploratory views.

In the first stage of a data science project is the acquisition of data. This process can occur in several different ways: 1. Generating semi-automated scenarios for real-time data collection and storing them with the auxilium of a Database Management System; 2. Using Integrated Business Management Systems so that indirectly the data is stored and made available to the scientist; 3. Through techniques of data extraction from the Web; 4. With the use of public databases; and many other possibilities.

know about new data analysis applications;
verify how the data analysis influences the decision making of an enterprise;
understand the results of using data analytics as business support;

upload, process and display data from the supplied dataset;
practice visual understanding of data;
generate visualization of data;
try to identify patterns (trends) in the analyzed database;
generate their own insights by relating and analyzing the data;

Think of a storytelling format to communicate the result.
Incluindo as análises abaixo:

Country with the largest movie center / series
Director with greater concentration of films / series
Categories with more titles
Perceptions generated from a time series
Relations between contents and their perceptions
Attributes of film and related series offerings
Tools and visualizations used to generate results
Perceptions and final interpretations generated with the analyses.

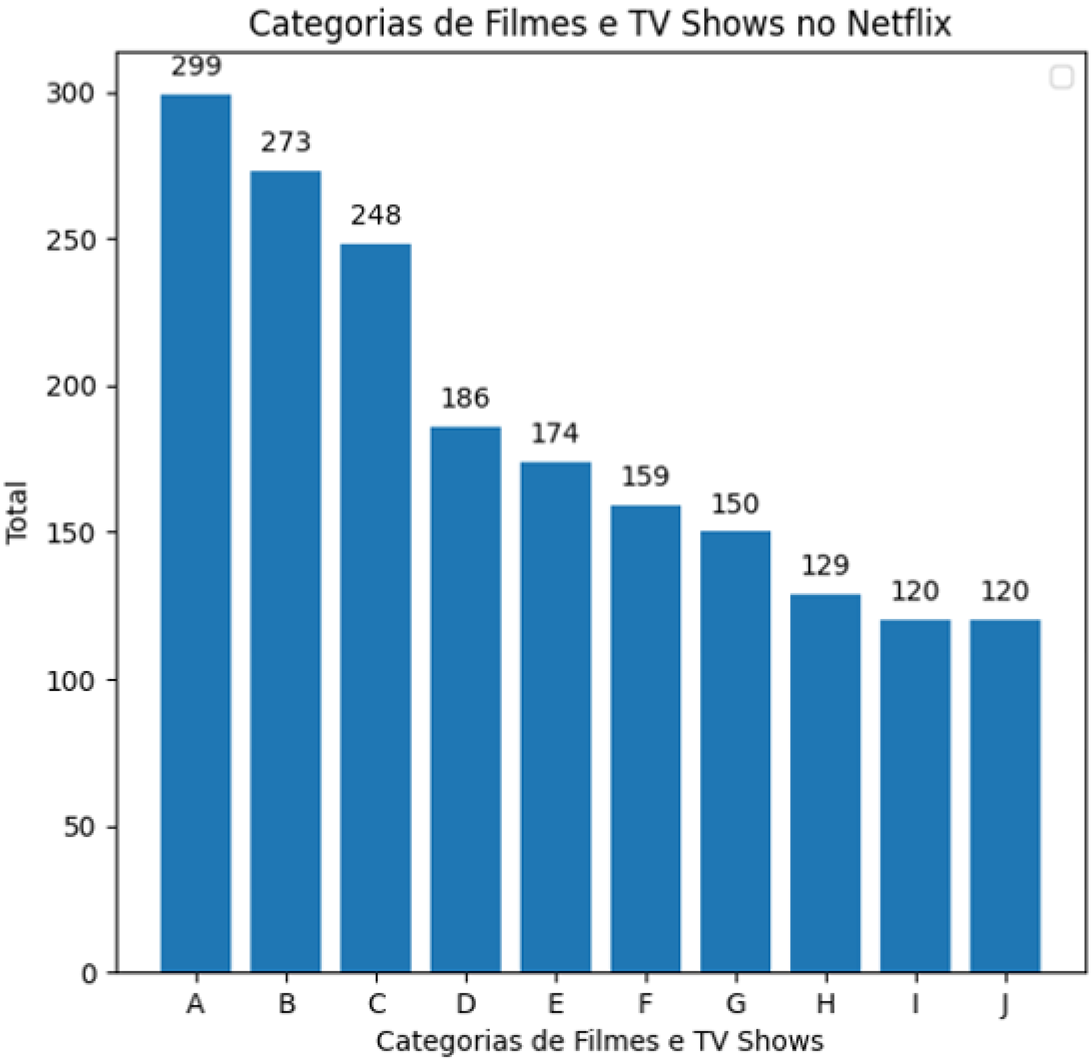
Phase 2:

In the second stage of the project can be defined as a stage of exploration, in which routines are executed that allow a better visualization of the data, without losing or compromising its essence. At this stage, we seek to represent the data in a way that is more easily analyzed. Usually segmented from the data, we test hypothesis, plot different types of graphs, we evaluate correlation between the variables studied and seek patterns and/ or variables that help us to build new perceptions. In this part of case study will deal with two steps that are fundamental for a (a) data scientist. Having the task of ensuring storage of this cloud data, using the IBMCloud platform with the DB2 database and structuring a sequence of instructions, documented in the format of a notebook, that allow you to interact and update this data and explore Netflix title data.
It will be able to query and return all records from your database; change the "rating" attribute of a particular entity informing the new classification and its ID; and consult and return all records in your database where the country and director are not null/empty.
See films released until 2015 and films released in the last 5 years.
Etapas:

- Directors with more catalogued films
- Directors with more catalogued series
- Years that had more releases
- Categories that are most represented in the Netflix catalog made available on this basis
- Countries that produced more securities released and registered on this basis
- Movies that are from the same franchise
- The average length of the cataloged film

- Impacts of storing your database in the cloud
- Graphics and histograms that support visualization
- Data visualization with Matplotlib.
- Access a database in the cloud using DB2
- SQL queries for access to data stored in the cloud
- Table responsible for storing movie information

1) Analysis of Movies and TV Shows by Categories – Netflix



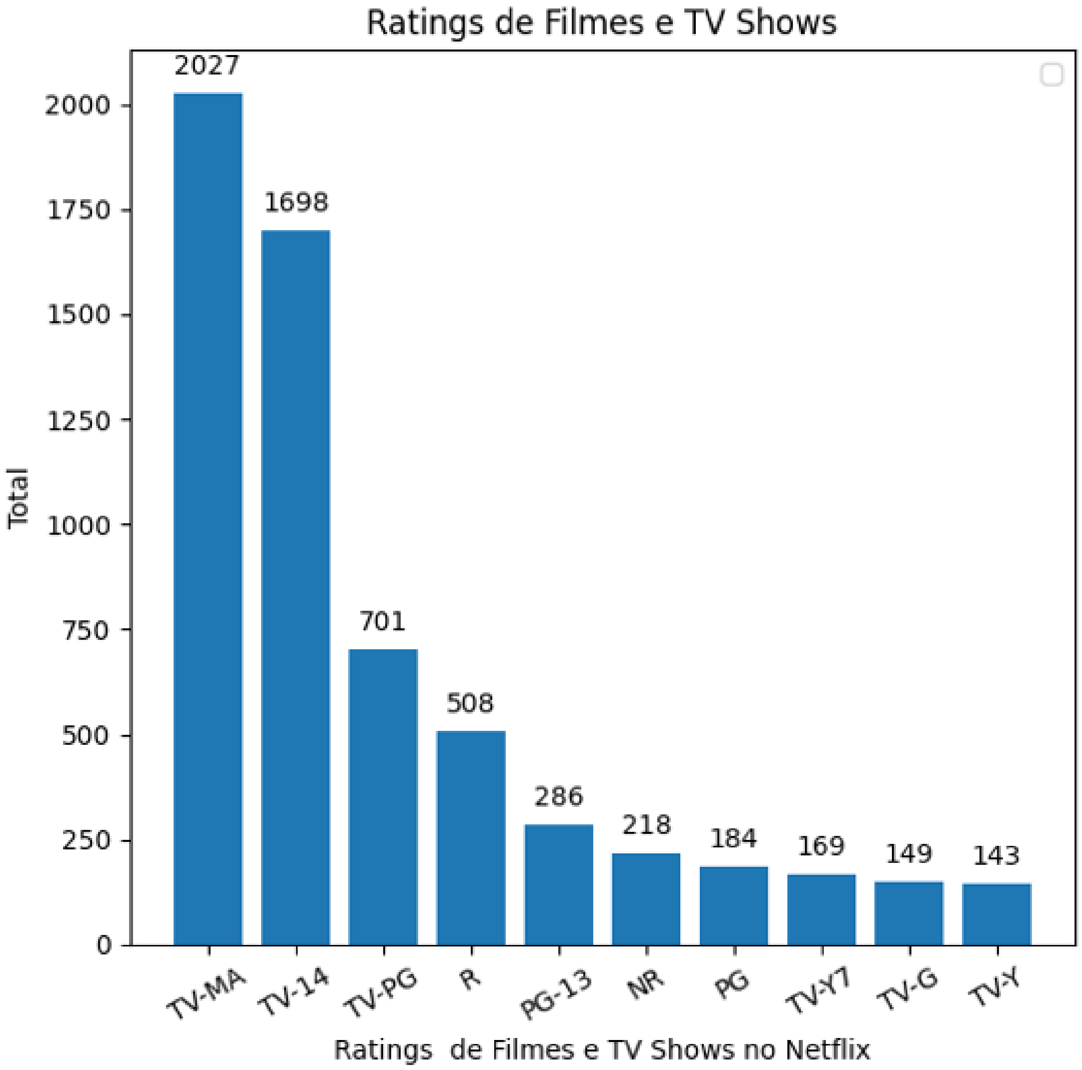
Legend

A	Documentaries
B	Stand-Up Comedy
C	Dramas, International
D	Dramas, Independent Movies, International Movies
E	Comedies, Dramas, International Movies
F	Kids' TV
G	Documentaries, International
H	Children & Family Movies, Comedies
I	Children & Family Movies

Code:

```
import matplotlib.pyplot as plt
import numpy as np
import csv

plt.rcdefaults()
fig, ax = plt.subplots()

x = []
y = []
with open('C:\\Temp\\project1.csv','r') as csvfile:
    plots = csv.reader(csvfile, delimiter='|')
    for row in plots:
        x.append(str(row[1]))
        y.append(int(row[0]))

plt.bar(x,y)
plt.title ('Categorias de Filmes e TV Shows no Netflix')
plt.xlabel('Categorias de Filmes e TV Shows')
plt.ylabel('Total')
plt.legend()

for p in ax.patches:
    ax.annotate(np.round(p.get_height(),decimals=2),
                (p.get_x()+p.get_width()/2., p.get_height()),
                ha='center',
                va='center',
                xytext=(0, 10),
                textcoords='offset points')
            plt.show()
```

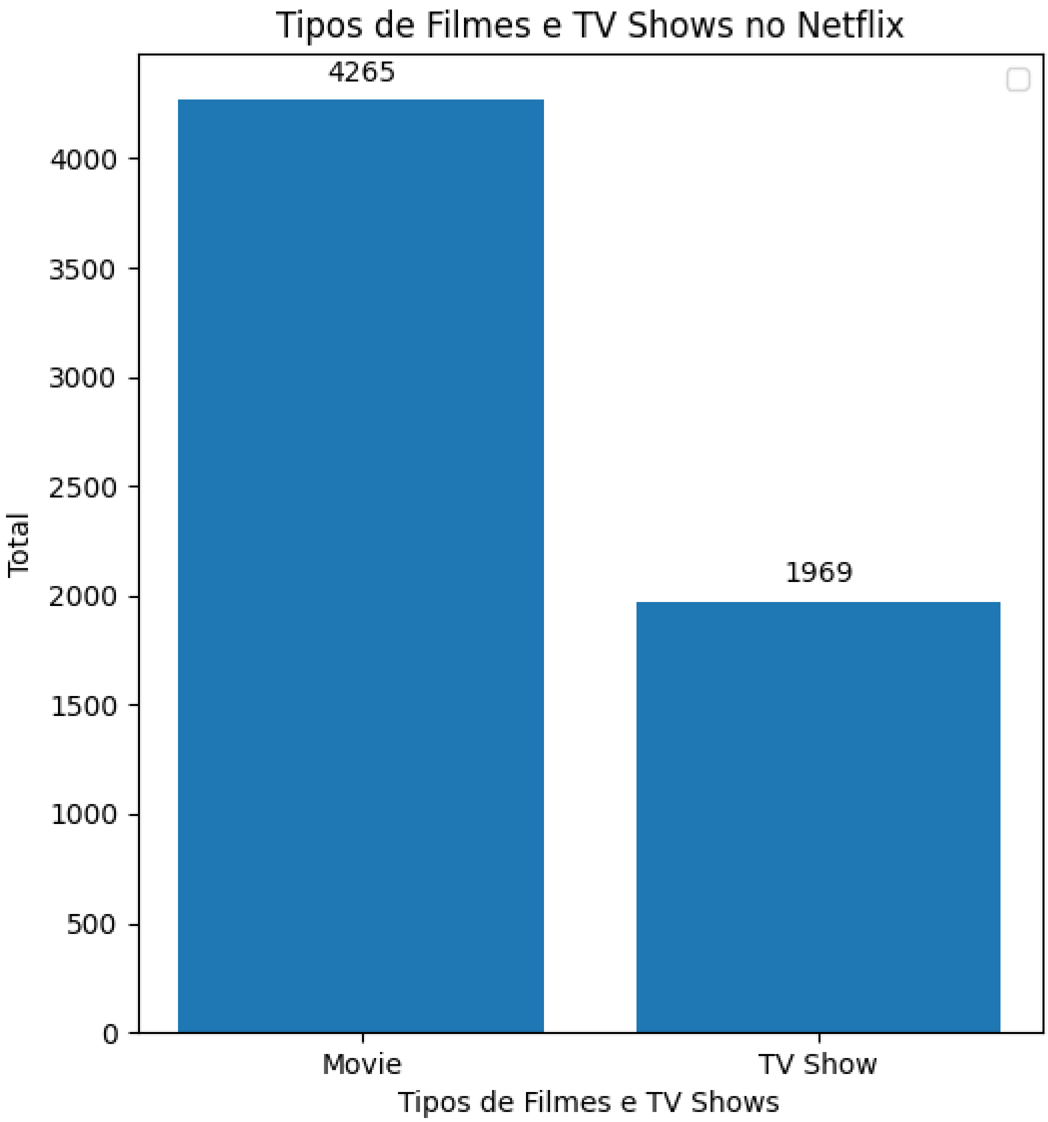
2) Análise de Filmes e TV Shows por Rating – Netflix



Code:

```
import matplotlib.pyplot as plt
import numpy as np
import csv


plt.rcdefaults()
fig, ax = plt.subplots()

x = []
y = []

with open('C:\\Temp\\project2.csv','r') as csvfile:
    plots = csv.reader(csvfile, delimiter=',')
    for row in plots:
        x.append(str(row[1]))
        y.append(int(row[0]))

plt.bar(x,y)
plt.xlabel('Ratings de Filmes e TV Shows no Netflix')
plt.ylabel('Total')
plt.legend()
plt.title ('Ratings de Filmes e TV Shows')
plt.xticks(rotation=30, horizontalalignment="center")

for p in ax.patches:
    ax.annotate(np.round(p.get_height(),decimals=2),
                (p.get_x()+p.get_width()/2., p.get_height()),
                ha='center',
                va='center',
                xytext=(0, 10),
                textcoords='offset points')

plt.show()
```

3) Analysis of Movies and TV Shows by Type - Netflix



Code:

```
import matplotlib.pyplot as plt
import numpy as np
import csv

plt.rcdefaults()
fig, ax = plt.subplots()

x = []
y = []

with open('C:\\Temp\\project3.csv','r') as csvfile:
    plots = csv.reader(csvfile, delimiter=',')
    for row in plots:
        x.append(str(row[1]))
        y.append(int(row[0]))

plt.bar(x,y)
plt.title ('Tipos de Filmes e TV Shows no Netflix')
plt.xlabel('Tipos de Filmes e TV Shows')
plt.ylabel('Total')
plt.legend()

for p in ax.patches:
    ax.annotate(np.round(p.get_height(),decimals=2),
                (p.get_x()+p.get_width()/2., p.get_height()),
                ha='center',
                va='center',
                xytext=(0, 10),
                textcoords='offset points')

plt.show()
```