**Final Project Report**

**Netflix Movies and TV Shows**

**Table of Contents**

[**1.** **Problem Statement** 3](#_Toc46907513)

[**2.** **DS Pipeline** 3](#_Toc46907514)

[**Pipeline** 3](#_Toc46907515)

[**3.** **Features and Data** 4](#_Toc46907516)

[**Dataset** 4](#_Toc46907517)

[**Data** 4](#_Toc46907518)

[**Features** 4](#_Toc46907519)

[**Reason to choose this dataset** 4](#_Toc46907520)

[**4.** **Data Pre-processing** 5](#_Toc46907521)

[**5.** **Exploratory Data Analysis (EDA)** 7](#_Toc46907522)

[**6.** **Modelling** 15](#_Toc46907523)

[**7.** **Explanation of the Findings** 21](#_Toc46907524)

# **Problem Statement**

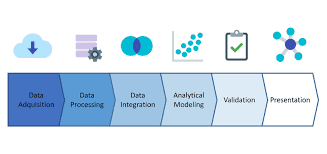
In last decades the use of online services are acquiring the market in all field. In few years the Netflix has captured huge market in field of online entertainment. It would be interesting to analyze how they have added the movies and TV shows to stay in competitive field. In order to predict the future behavior. We have built a predictive model which will predict the worldwide popularity of Movies and TV shows. We have trained our model using regression technique.

# **DS Pipeline**

A **data science pipeline** is the overall step by step process towards obtaining, cleaning, visualizing, modelling, and interpreting **data** within a business or group. **Data science pipelines** work for small **data** analysis.

## **Pipeline**

The following image shows how the dataset pipeline is followed for this process. And will be observed during the whole documentation.



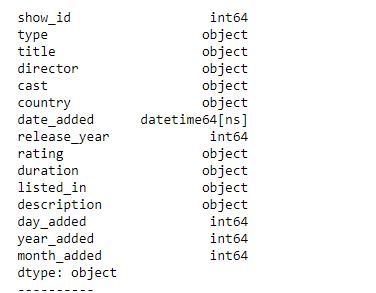
# **Features and Data**

## **Dataset**

We have selected Netflix Movies and TV Shows dataset which have record of different Movies and TV shows. The dataset has more than 10000 entries and it has the following attributes along with datatypes.

## **Data**

The dataset includes the following columns.

****

## **Features**

Features of the data are as follows:

* **Show\_id:** It is the id of records.
* **Type:** Type tells whether it is Movie or TV Show.
* **Title:** Title is the name of Movie and TV Show.
* **Cast:** Cast is name of people acting in the Movie or TV Show.
* **Date added:** It is the release date of Movies and Shows.
* **Release \_year:** Release year contains only year of release.
* **Rating:** Contains ratings of Movies and Shows.
* **Duration:** It contains time duration of Movies and Shows.
* **Listed\_in:** It contains categories of viewers.
* **Description:** It contains small story line of Movies and Shows.

## **Reason to choose this dataset**

1. This was the most detailed dataset that we have searched so far about Movies and TV Shows. Plus we have very less missing values in this dataset
2. The data set was **.csv** file

# **Data Pre-processing**

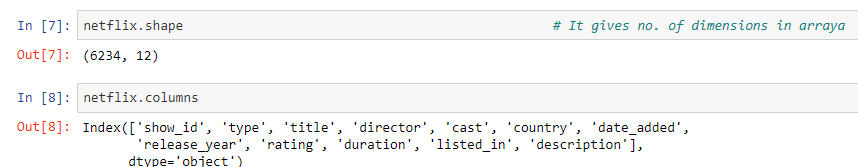
1. **Import Python libraries**



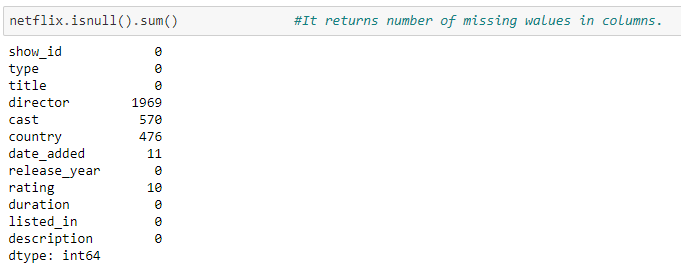
1. **Read csv**



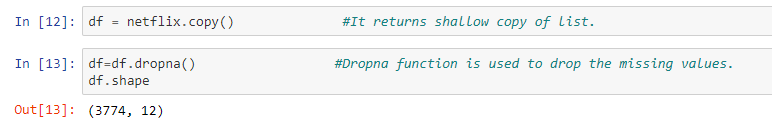
1. **Checking Columns and Shapes**



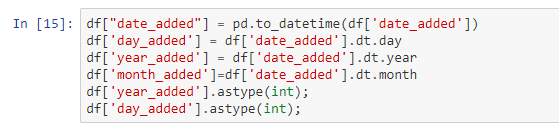
1. **Null Values Check**



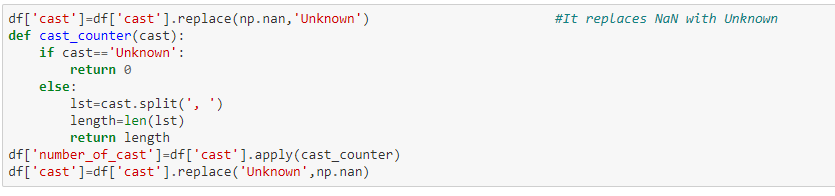
1. **Creating shallow copy and Dropping null values**



1. **Changing Date Pattern**



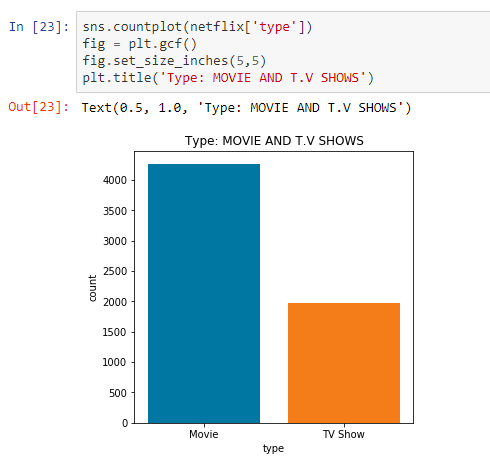
1. **Replacing with unknown**



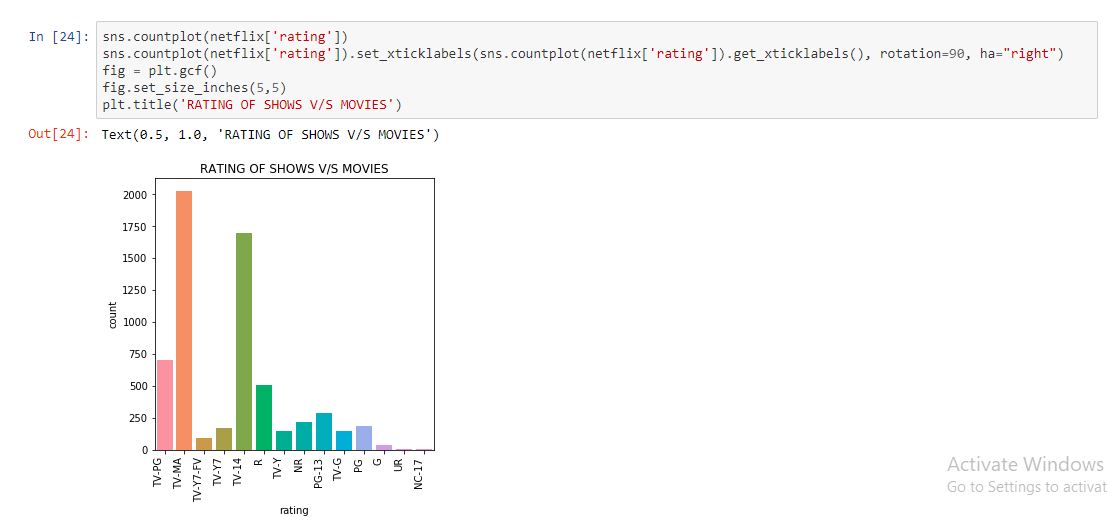
# **Exploratory Data Analysis (EDA)**

By now our data was cleaned and then we performed EDA on it.

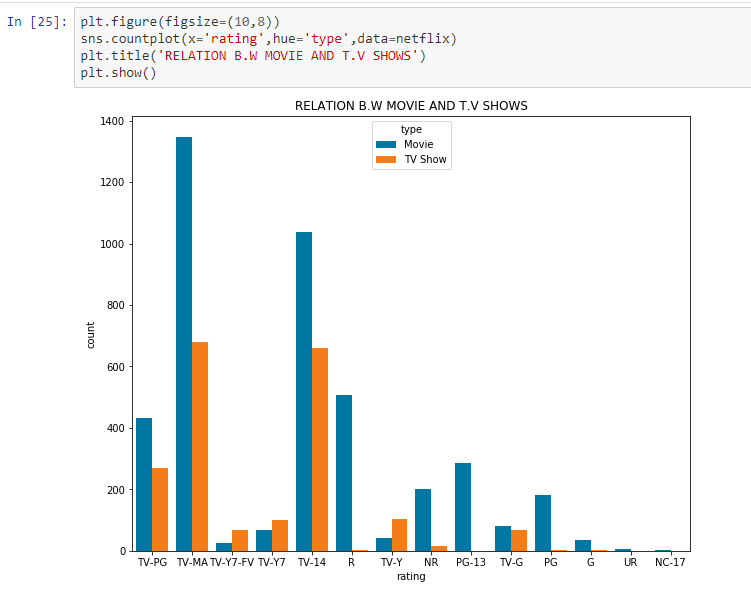
1. **Plot Graph for Type**



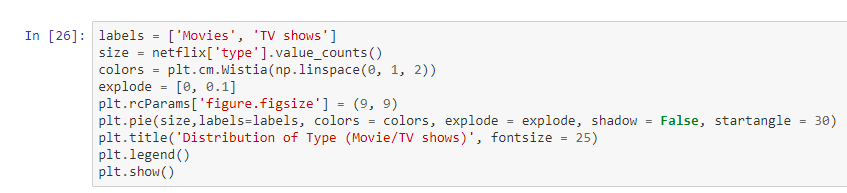
1. **Ratings Graphical Representation**

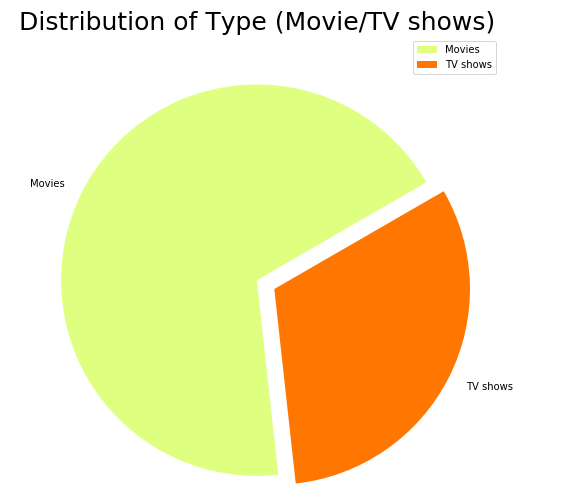


1. **Relation between Movies and TV Shows**

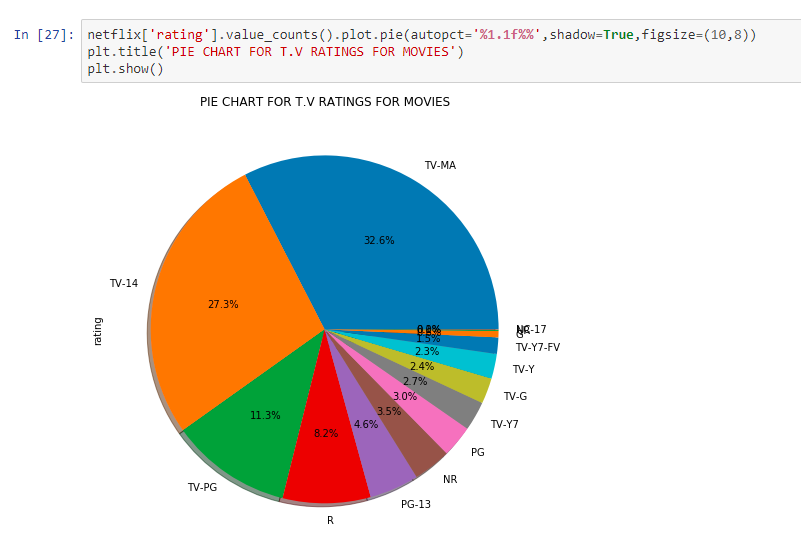


1. **Distribution of Movies and TV Shows**

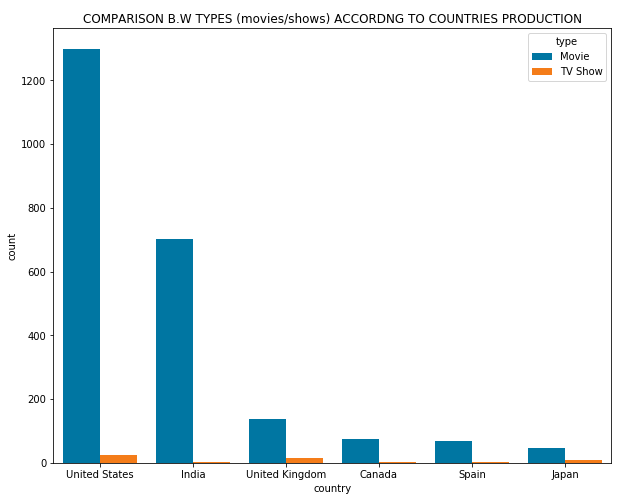
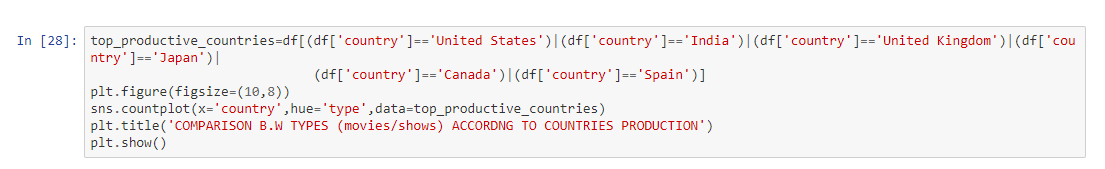




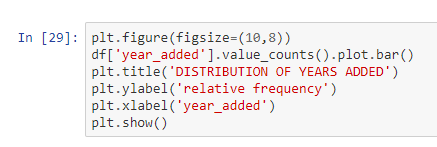
1. **Pie Chart of Ratings of Movies and TV Shows**

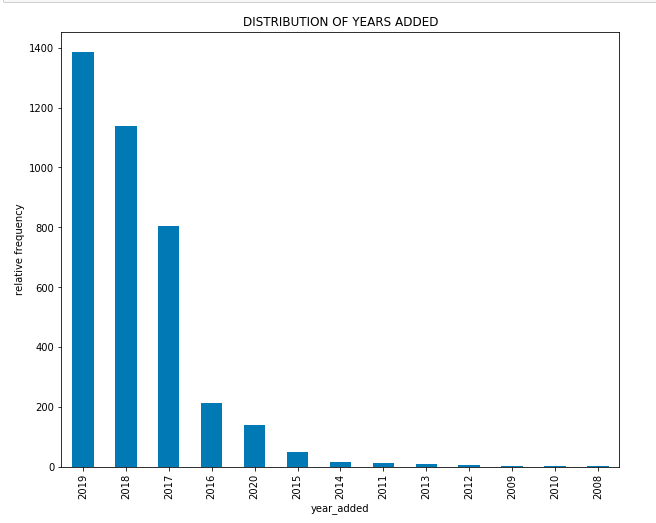


1. **Comparison according to Countries**

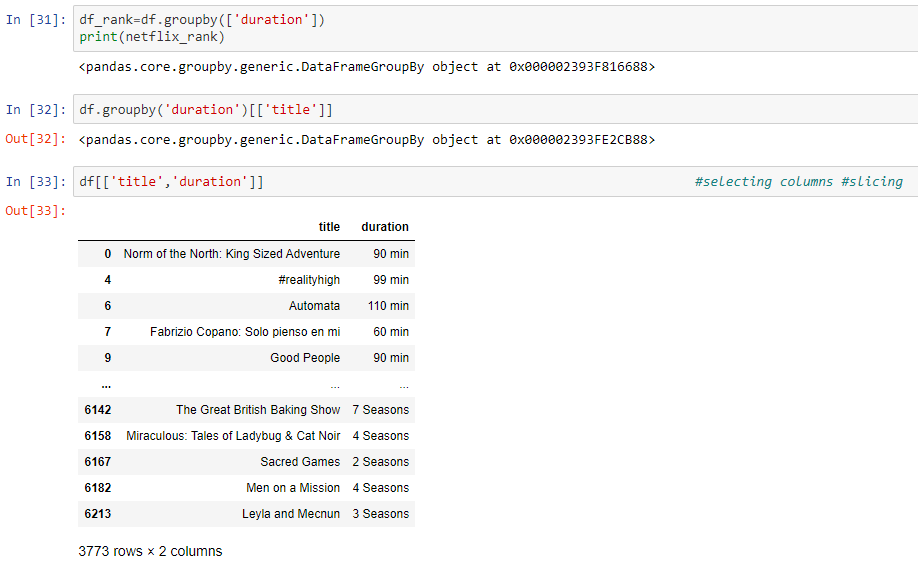


1. **Distribution of years added**

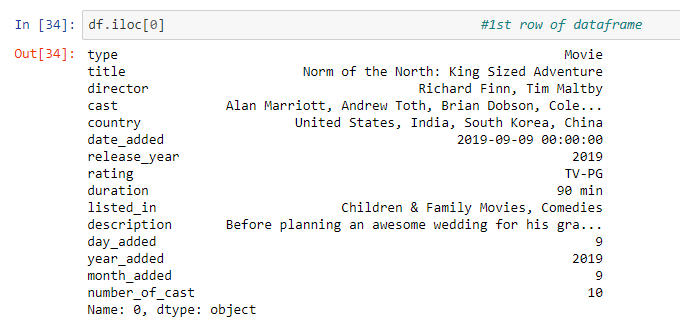




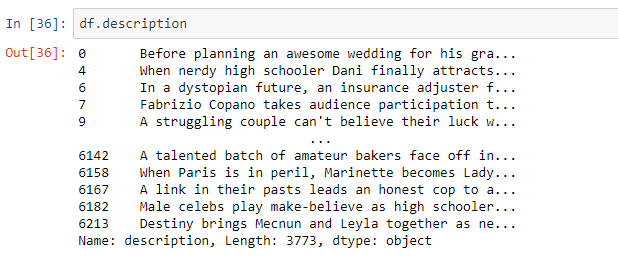
1. **Slicing**



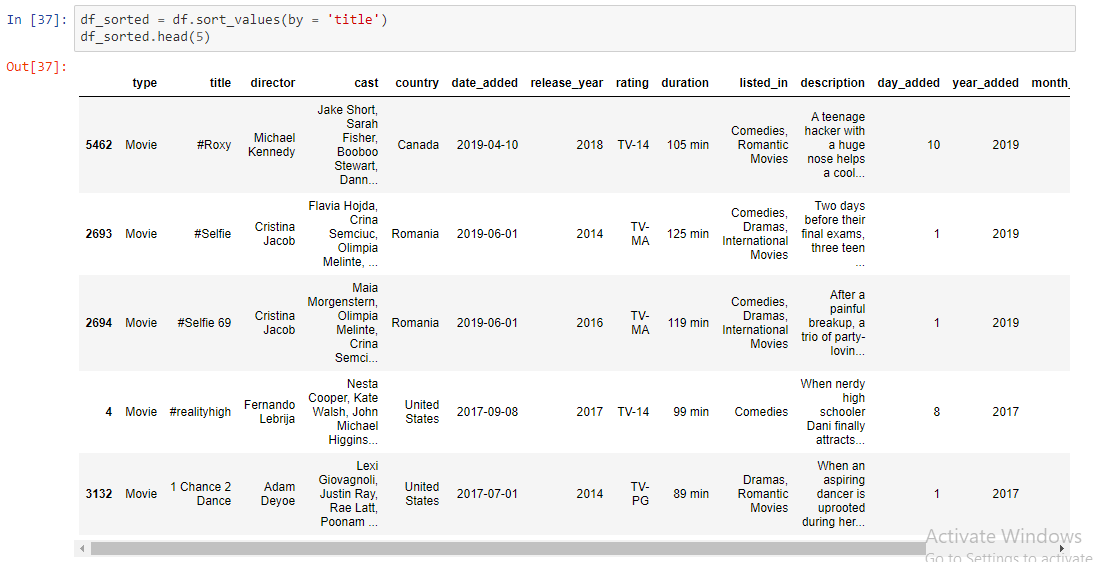
1. **I-loc**



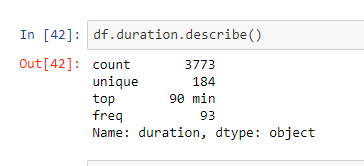
1. **Description**



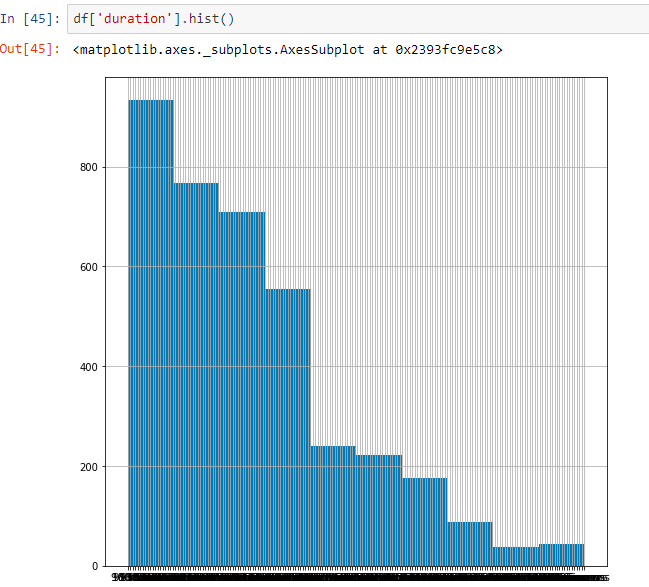
1. **Sorting**



1. **Duration**



1. **Histogram for Duration**

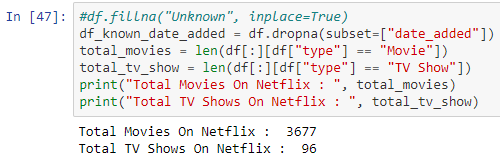


# **Modelling**

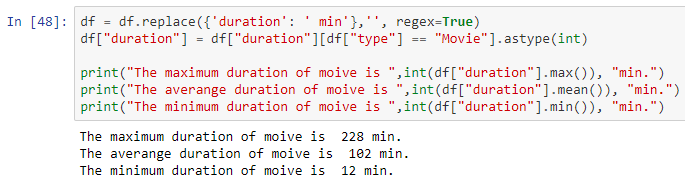
1. Regression technique is used for the dataset modelling.
2. The main reason why this model is selected is because it indicates the significant relationships between dependent variable and independent variable.
3. This technique proved effective to a big extent.
4. Classification technique can also be used on the given dataset.
5. **Import library for model**



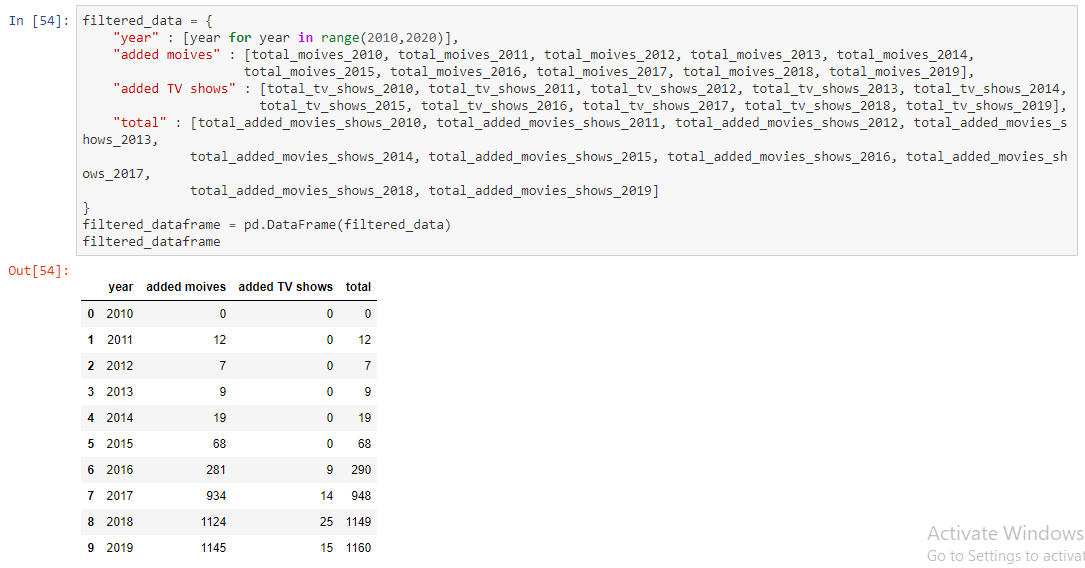
1. **Count of Movies and Shows**



1. **Duration Count**

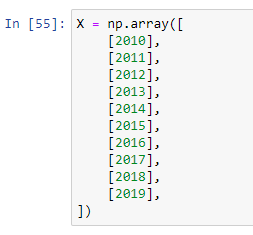


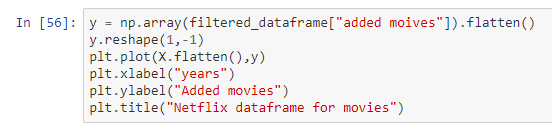
1. **Filter Data for Modelling**



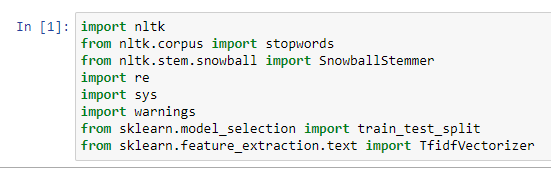
1. **Data frame for Movies Prediction**

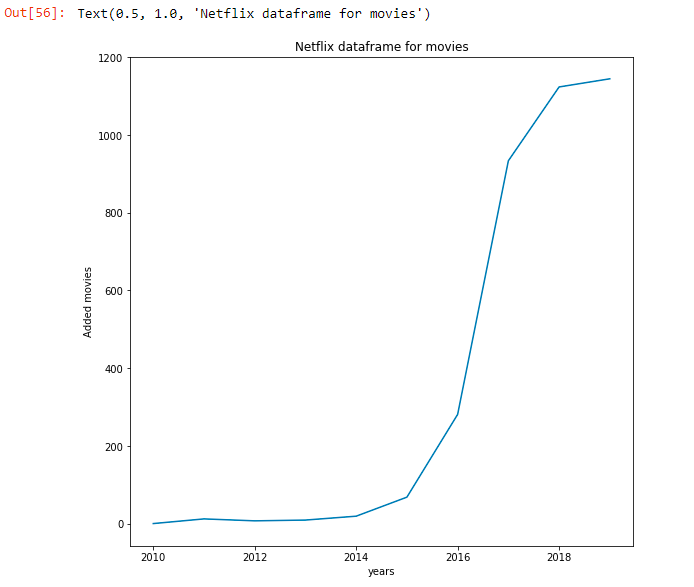
Now, we are going to try the linear regression model. Let’s start by selecting the independent features and storing them in a variable named X, and our independent feature in a variable named y.



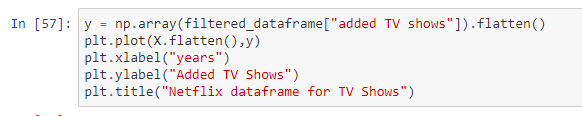


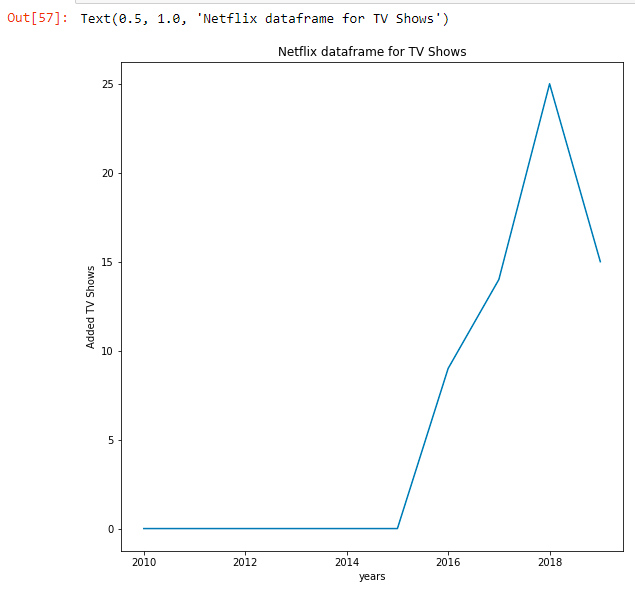
**Libraries for Splitting into test and train Data**



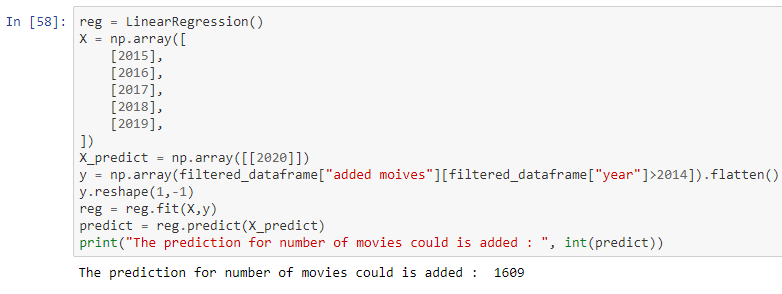


1. **Data frame for TV Shows Predictions**

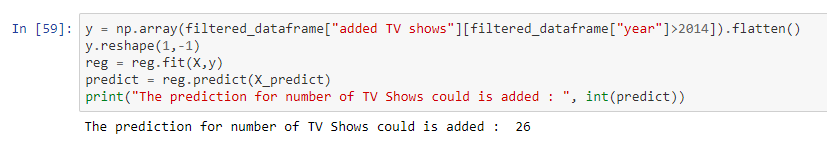




1. **Prediction for Movies**



1. **Prediction for TV Shows**



# **Explanation of the Findings**

The above project is done using Jupyter Notebook with help of python language. By observing the graph above it is wise to analyze the last five years data (Year: 2015-2019). **Regression model** from sklearn library is used and data is also split into test and training data. We have done preprocessing, EDA and Modelling using the same pipeline.

The predictions are clearly showing that the trend of movies in increasing and trend of watching TV Shows is decreasing.

Without going into detail of Linear Regression model let's jump to final predictions.

* **The prediction for number of movies could be added in 2020. : 2014**
* **The prediction for number of TV Shows could be added in 2020 : 933**