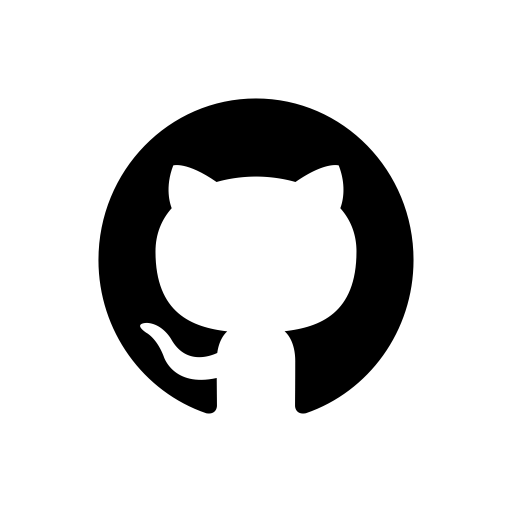
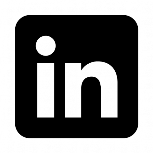
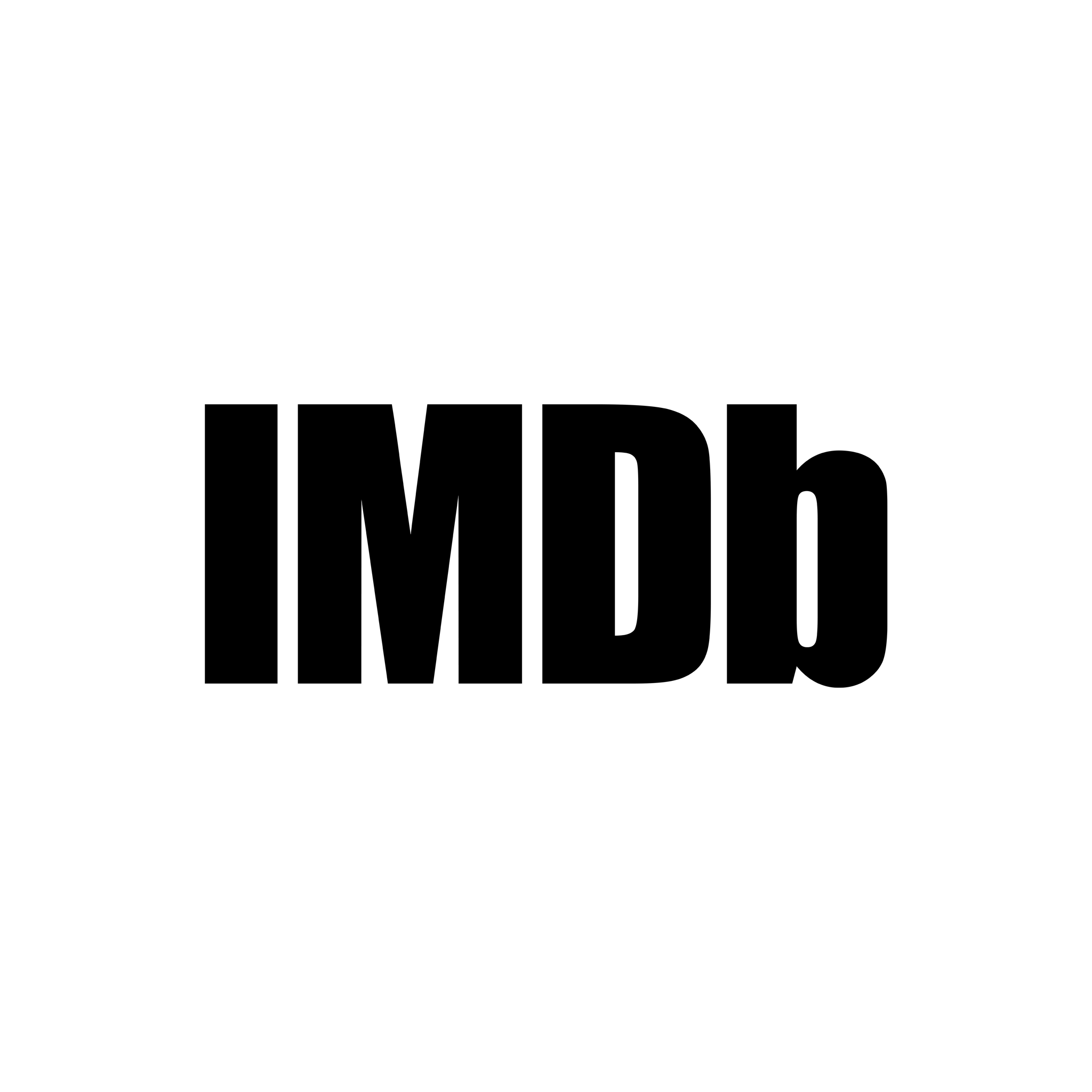
[](https://github.com/sidjaku)[](https://www.linkedin.com/in/sidjaku/)

Unveiling cinematic treasures, we've swept away the clutter from movie descriptions while counting every vote without a comma in sight. Embarking on a data odyssey, our "ONE-LINE" column undergoes a cinematic makeover, shedding special characters and 'Add a Plot' musings to reveal concise movie summaries. Meanwhile, our "VOTES" column emerges from a numerical cleanse, bidding adieu to commas and embracing pristine numeric clarity. As for our "YEAR" column, it's a whimsical expedition through entertainment epochs, where even the humble "nan" and enigmatic parentheses add a touch of intrigue to the storytelling tapestry. Lights, camera, clean data - our project is ready! for its blockbuster debut!

DATA CLEANING

Working with Movies and TV Shows Dataset!

Contents

[Movies and TV Shows Dataset Analysis 3](#_Toc164054068)

[Overview 3](#_Toc164054069)

[Dataset Description 3](#_Toc164054070)

[Data Cleaning and Preprocessing 3](#_Toc164054071)

[Key Findings 3](#_Toc164054072)

[Before 4](#_Toc164054073)

[After 4](#_Toc164054074)

[The Year Column (Numeric) 5](#_Toc164054075)

[The Genre column (String) 6](#_Toc164054076)

[The One-Line column (String) 7](#_Toc164054077)

[The Votes column (Numeric) 8](#_Toc164054078)

[The Gross column (lol) 8](#_Toc164054079)

[The Stars Column (String-Descriptive) 9](#_Toc164054080)

## Movies and TV Shows Dataset Analysis

### Overview

This project aims to analyse a dataset containing information about various movies and TV shows. The dataset includes several columns such as 'Movies', 'Year', 'Genre', 'Rating', 'One-Line', 'Stars', 'Votes', 'RunTime', and 'Gross'. The data is cleaned and processed to derive insights and patterns from the entertainment industry.

### Dataset Description

* **Movies**: Name of the Movie/TV Show.
* **Year**: Year of the release.
* **Genre**: Genre of the Movie/Show.
* **Rating**: Movie/Show Rating.
* **One-Line**: Concise description or summary of the Movie/Show.
* **Stars**: Contains information about the cast members and, in some cases, details about directors.
* **Votes**: Total number of votes received for the Movie/Show.
* **RunTime**: Duration or runtime of the Movie/Show.
* **Gross**: Total Amount earned Worldwide.

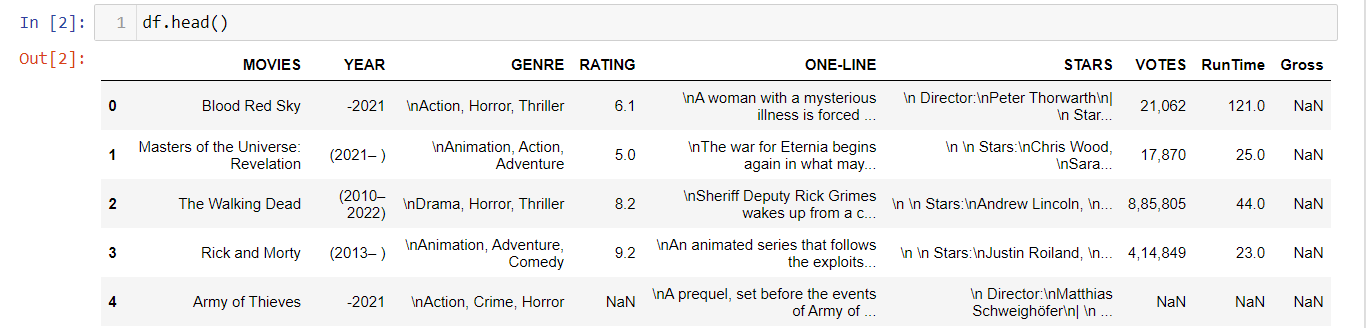
### Data Cleaning and Preprocessing

* **Handling Missing Values**: Addressed missing values in columns where applicable.
* **Parsing Year and Runtime**: Extracted and converted year and runtime information into appropriate formats for analysis.
* **Cleaning 'Stars' Column**: Extracted cast member names and handled cases with additional director information.

### Key Findings

* Data Cleaning is mandatory before making analysis. It is common to have unorganized data especially where manual input is involved.
* Using regex for cleaning is a good method. It is a must learn library in python if you are into analysis.
* Finding where to use brute force is crucial as well. For this dataset, I had to analyse it by length since using normal regex wouldn’t work.
* Data cleaning is a time-consuming process if you are new to it. It requires finding patterns and knowing what logic to apply to make sense out of the data. For ex- You can find the year column having various 0 values, we need to create a separate DataFrame without these 0 values.
* Handling missing data depends on what data you are using. There is no point in adding the mean of the year columns to the missing year values, because it does not make sense.

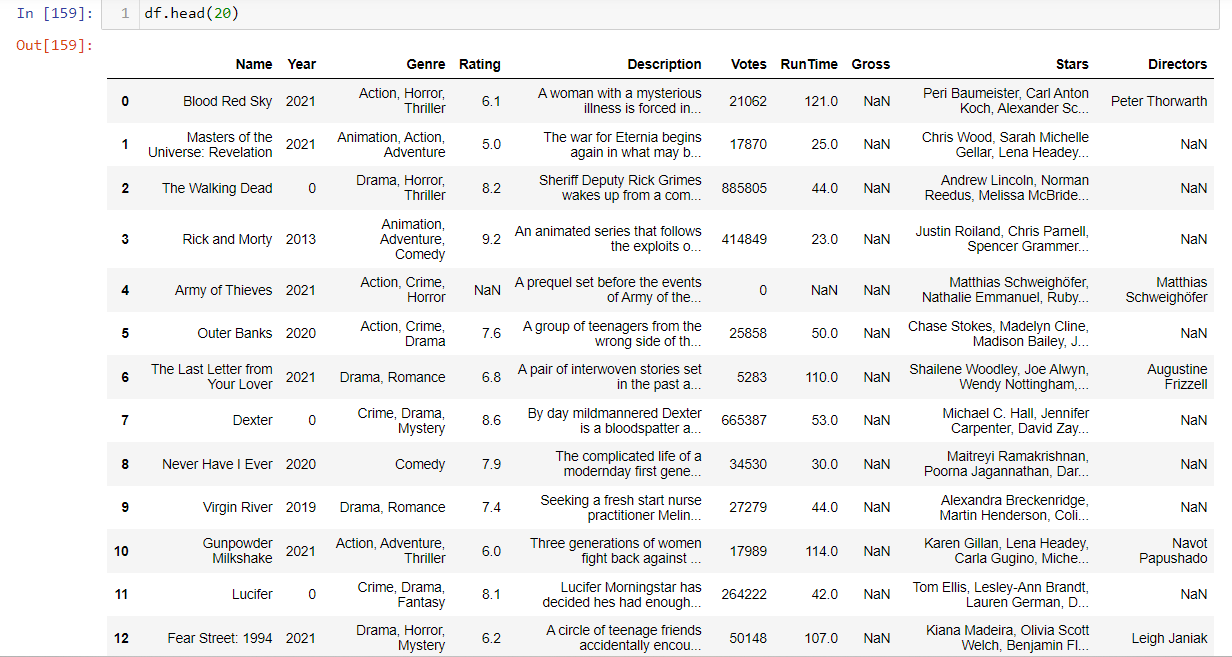
## Before

This is a glimpse of the given dataset!

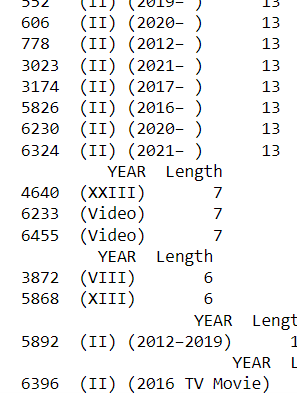
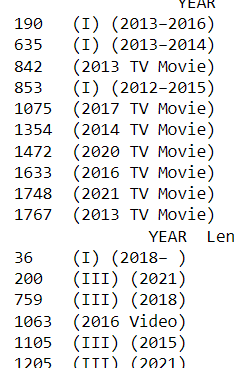
* Apart from the Movies column, all the other columns needed cleaning in order to perform a suitable analysis.
* The Year column has multiple anomalies.
* The Genre, One-Line and Stars have ‘\n’ at various places, suggesting that it might have been parsed from a CSV file.
* The Stars column has both Directors and Stars in them.
* The Votes column needs to be in an integer datatype.

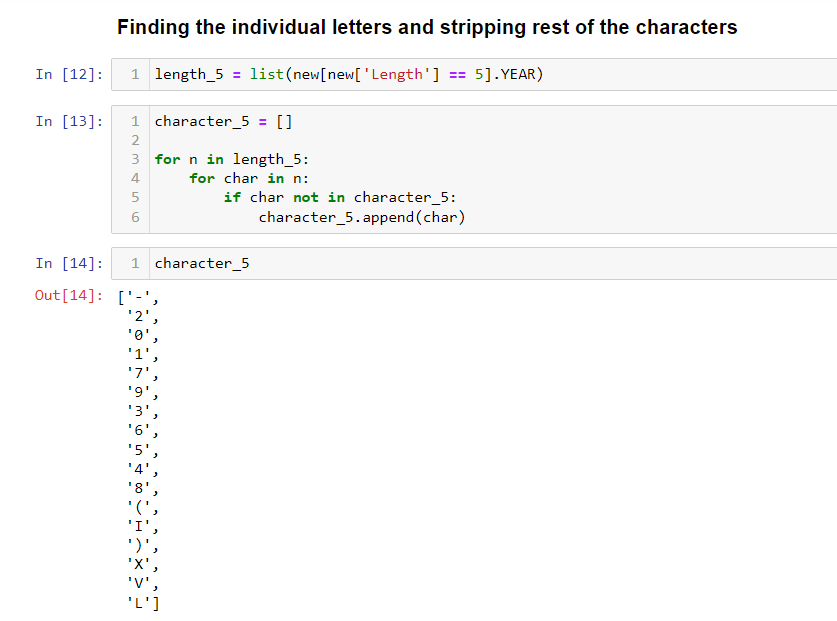
Let’s go with these columns 1 by 1!

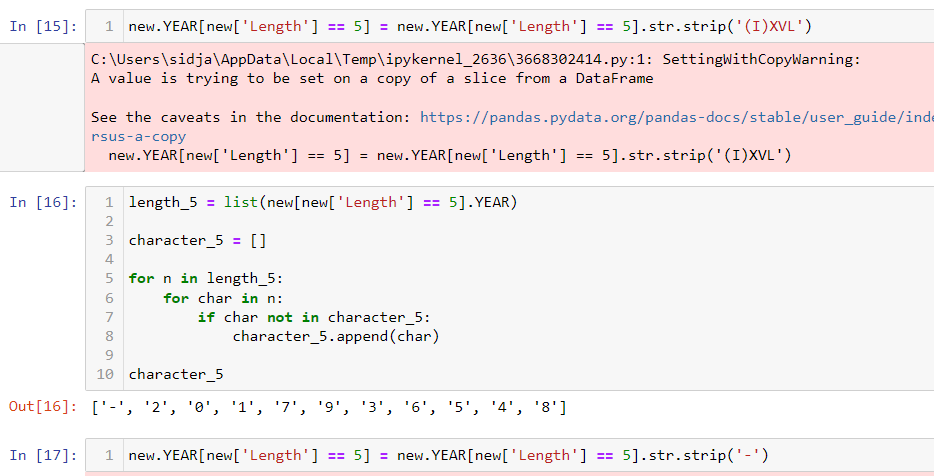
## After



## The Year Column (Numeric)

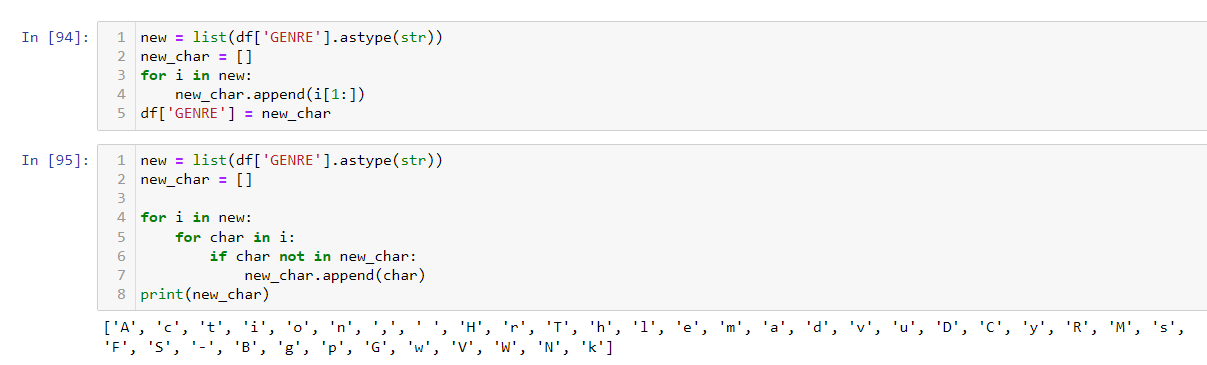
As you can see, I took the approach of cleaning the year column by analysing its length. I iterated the characters of each length and thankfully I could find a similar pattern in them. For example – the characters having 5 as length-



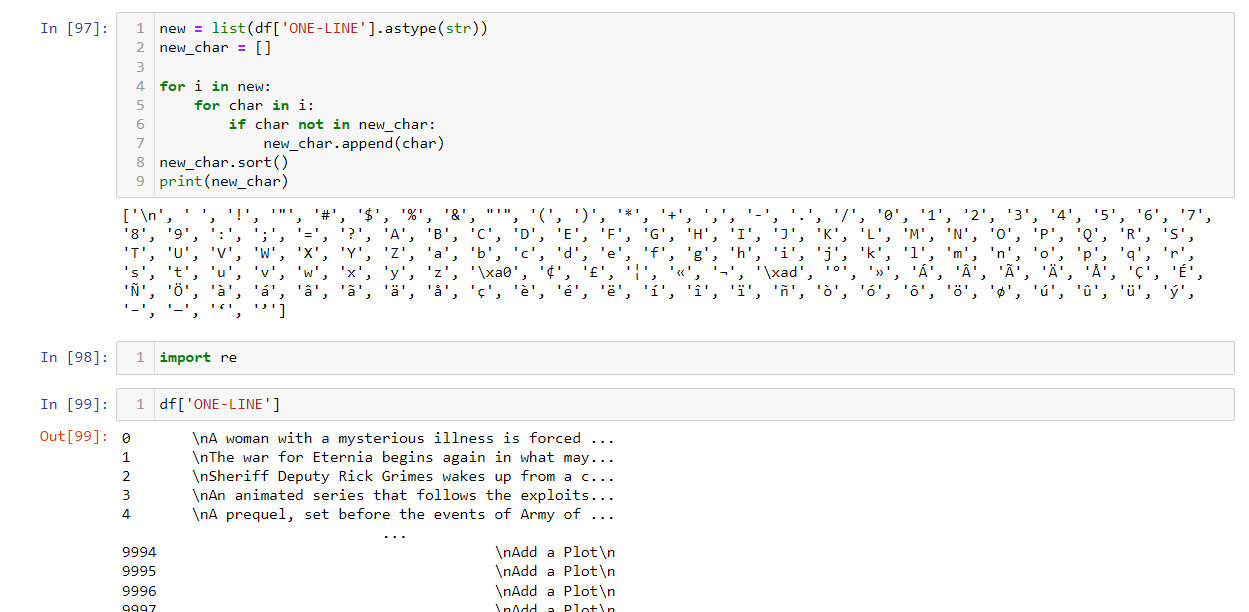


This was by far the messiest column to clean. It is evident that these were manually inputted at many cases. This would not be the case in today’s time where many data points are dropdown or click based to mitigate these manual errors.

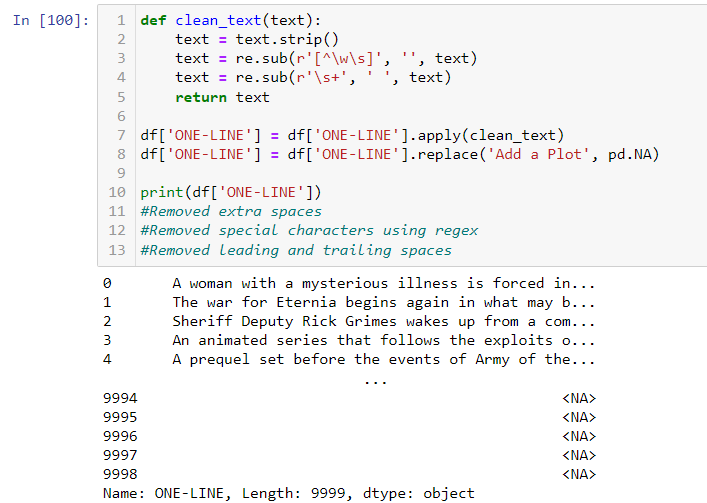
## The Genre column (String)



## The One-Line column (String)

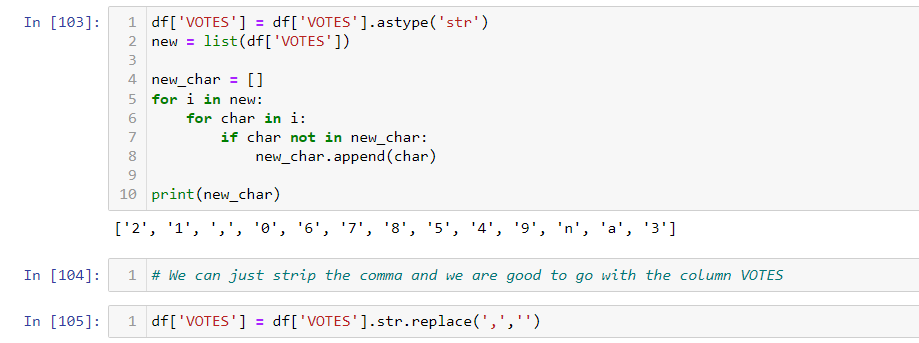
This is what the One-Line column looks like – 

It has a lot of random characters. Well, not really random since they might have been written using different characters of regional languages. But they needed to be cleaned never-the-less.  
I created a simple function using regex to clean this column

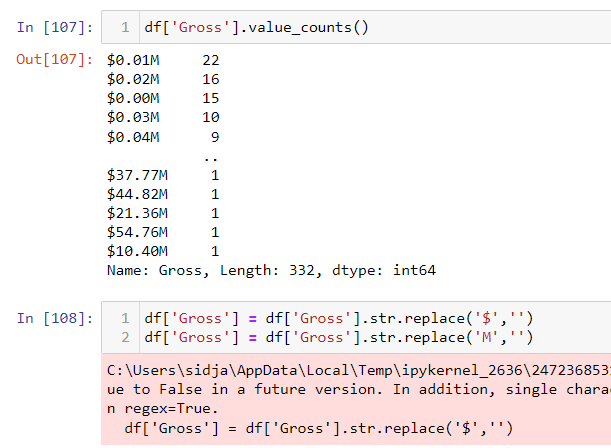


## The Votes column (Numeric)

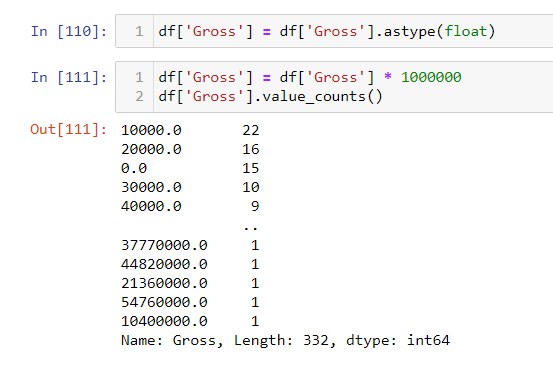
For the vote column, I just needed to strip the comma.



## The Gross column (lol)

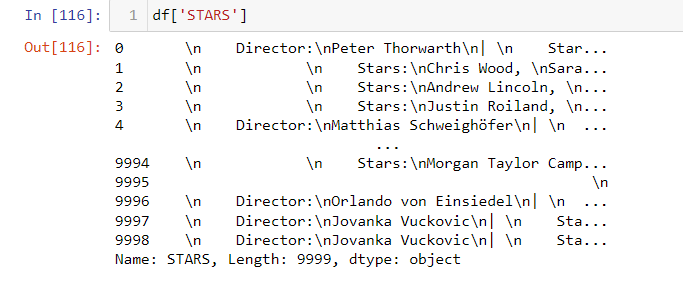


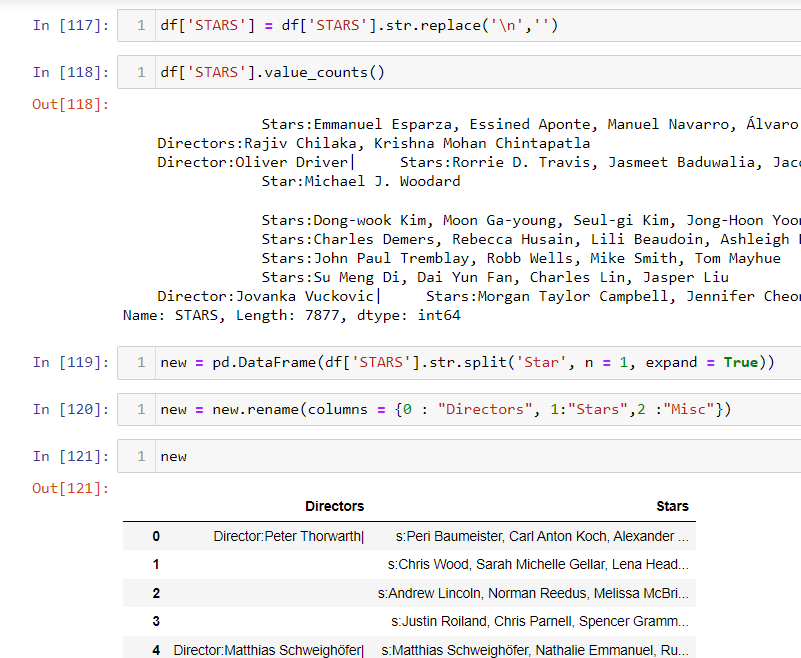
Just replaced the $ and M and multiplied the value by a million.



## The Stars Column (String-Descriptive)

Nothing makes sense here.

Decided to split the column after replacing the ‘\n’ values.



Then I had to replace the ‘|’ characters. Then I analysed by the length and the characters present as well.

