**IMDB Dataset Data Cleaning Plan**

*There are 13 columns named are –*

1. *Movie name*
2. *Year of release*
3. *Watch hour*
4. *Rating*
5. *Ratedby*
6. *Film Industry*
7. *Genre*
8. *Director*
9. *Box office collection*
10. *User reviews*
11. *Awards*
12. *Description*
13. *Streaming platform*

As we know that there are 1 column which is float64 datatype , 2 columns has int64 datatype and 10 columns have object or string datatype .

|  |  |
| --- | --- |
| Movie name | object |
| Watch hour | object |
| Ratedby | object |
| Film Industry | object |
| Genre | object |
| Director | object |
| Box office collection | Object |
| Awards | object |
| Description | object |
| Streaming platform | object |

|  |
| --- |
| Year of release int64 |
| Rating float64 |
| User reviews int64 |

* Memory usages by the dataset is 25.5 kb+
* There is some null values in our dataset .
* No duplicated values in our imdb dataset .
* In this dataset we have 250 rows and 12 columns .

Process Which we can follow :-

1. **Columns Which are not used =>**

We Don’t use Movie name, Description , Genre, Director . So we can either drop these columns after creating the copy of the Data- Frame or we don’t use these columns in any part of my analysing .

1. **Removing the Null Values =>**

As we see that in our dataset the column name Box Office Collection has some values like NIL which we consider as a Null Values . So Firstly we can check the percentage of the Null or Missing values of w. r. t. column and if the missing values has above the 5 % then we can fill with Mean , or Median , or Mode , or Randomly a constant variable . And if the missing values has less than 5 % of our column then without any hesitation we can remove / drop the Missing Values by using dropna – function.

1. **Data Cleaning =>**

We can clean the data like in Box Office Collection we can remove “$” sign and Rated by column we can remove “ k “character and many more .

1. **Columns in which we change the datatype =>**

We can change the datatype of Columns like Watch hour , Rated by , Box Office Collection , awards .

And also we can reduce the memory of the dataset by changing the datatype of all the columns and making it int32 datatype .

1. **Handling the Outliers =>**

After all the portion we can check the outliers and if we found that there is a outlier in our dataset then either we can remove the outlier by droping or we can handle the outlier by using Inter Quantile Range .

1. **Capture the insights =>**

After handling the outliers we can capture the insights by using graphs . We can use bar chart , pie chart , combo chart and many more to capture the insights of the dataset .

1. **Normalisation =>**

After handling the outliers and null values we can check the Normalization curve of the dataset . And if the curve is not Normally distributed then we can Normalise the complete dataset or just columns and make all the columns into a single range .

1. **Correlation =>**

Check out the Correlation function and choose those columns whose are really valuable for us .

1. **Conclusion =>**

If all the things we can done we can write a simple conclusion which I found of the dataset .