**Report Predict Movie Success**

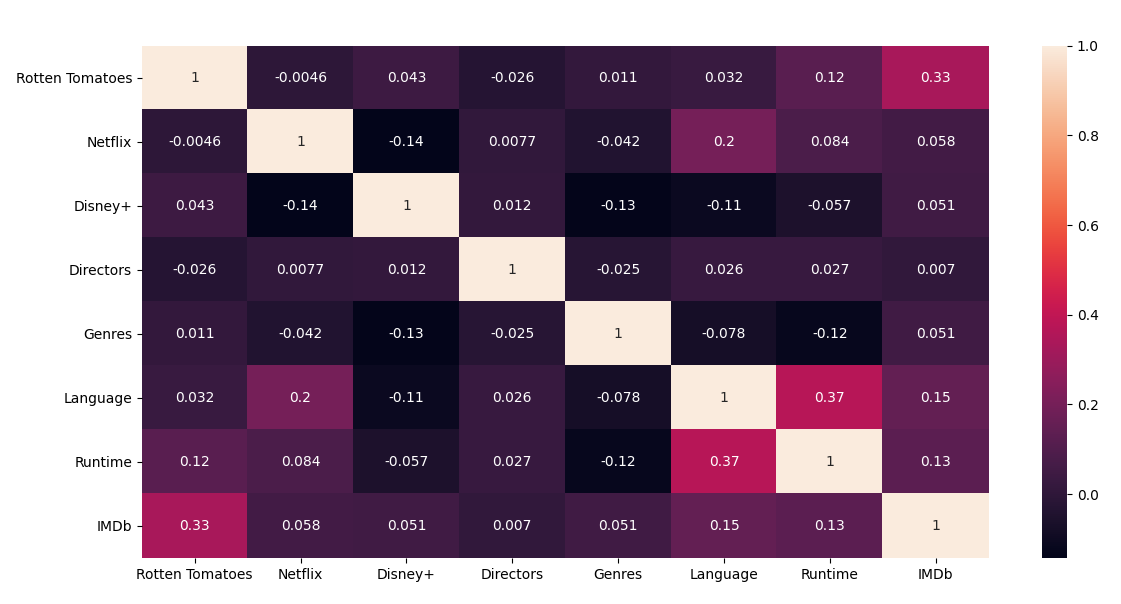
1. preprocessing techniques

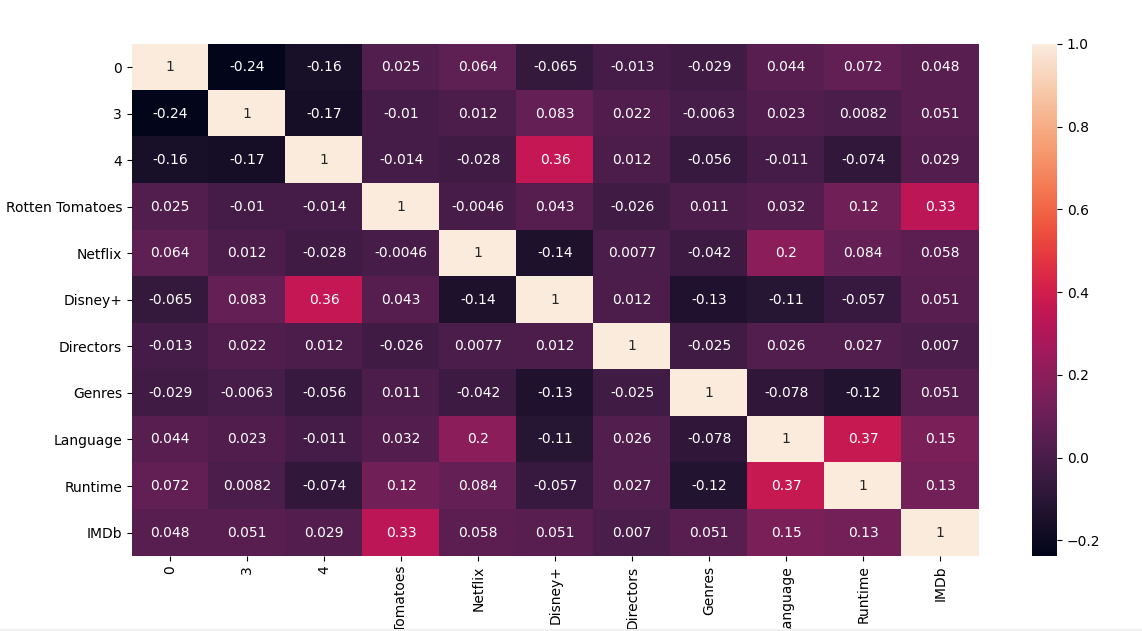
* solve missing data by:
* interpolate for (Rotten Tomatoes, age).
* encoding for categorical data (6 columns):

['Directors', 'Genres', 'Country', 'Rotten Tomatoes', 'Language',’age’]

* OneHotEncoder for age 🡪the least one of different values.
* OrdinalEncoder for the rest .
* Normalization.
* Drop columns that does not affected in result
* Split data for train and test.

1. Correlation:





1. **Regression techniques:**

* **Multi Linear Regression:** It is the basic and commonly used type for predictive analysis. It is a statistical approach to modelling the relationship between a dependent variable and a given set of independent variables. Multiple Linear Regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data.

Y = b0 + b1 \* x1 + b2 \* x2 + b3 \* x3 + …… bn \* xn

Y = Dependent variable and x1, x2, x3, …… xn = multiple independent variables

Step #1: Data Pre-Processing

Step #2: Fitting Multiple Linear Regression to the Training set

Step #3: Predicting the Test set results.

* **Polynomial Regression:** (second degree) Polynomial Regression is a powerful technique to encounter the situations where a quadratic, cubic or a higher degree nonlinear relationship exists.

Y = b0 + b1 \* x + b2 \* x2 + b3 \* x3 + …… bn \* xn

1. **differences between each model:**

* Mean Square Error:

Multi Linear Regression: 0.019696934664050706

Polynomial: 0.017397115586387072

* Time for training:

Multi Linear Regression: 10 sec.

Polynomial: 7 sec.

1. **Features in models:**

Title has been dropped because it has the least correlation with IMDb.

Used:

['year','age', 'Rotten Tomatoes', 'Netflix', 'Hulu', 'Prime Video', 'Disney+', 'Type', 'Directors', 'Genres','Country', 'Language', 'Runtime'].

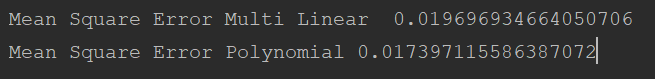
1. Size of Data:

* train 7419
* test 3180

1. further techniques:

* OneHotEncoder.
* OrdinalEncoder.
* Remove missing values.

1. Screen shoot:



1. Conclusion:

When use onehotencoding(‘age’)

Polynomial at 2 degree is better than multi linear in error,

Multi Linear Regression: 0.019696934664050706

Polynomial: 0.017397115586387072

When use onehotencoding('Rotten Tomatoes')

Multi Linear Regression: 0.016

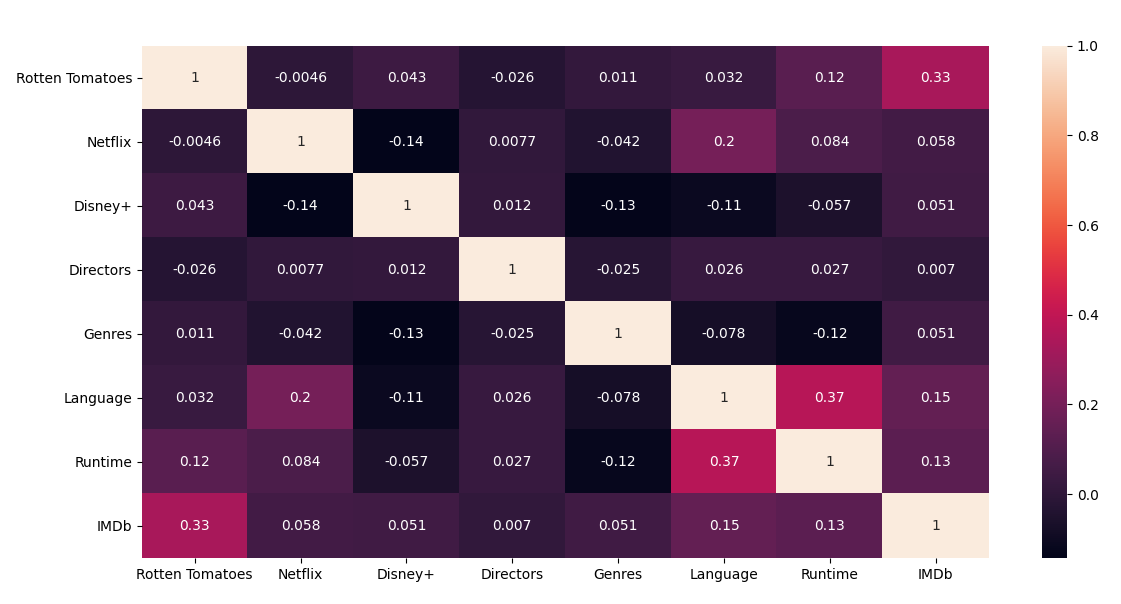
Polynomial: size of columns is too large and not calculated.

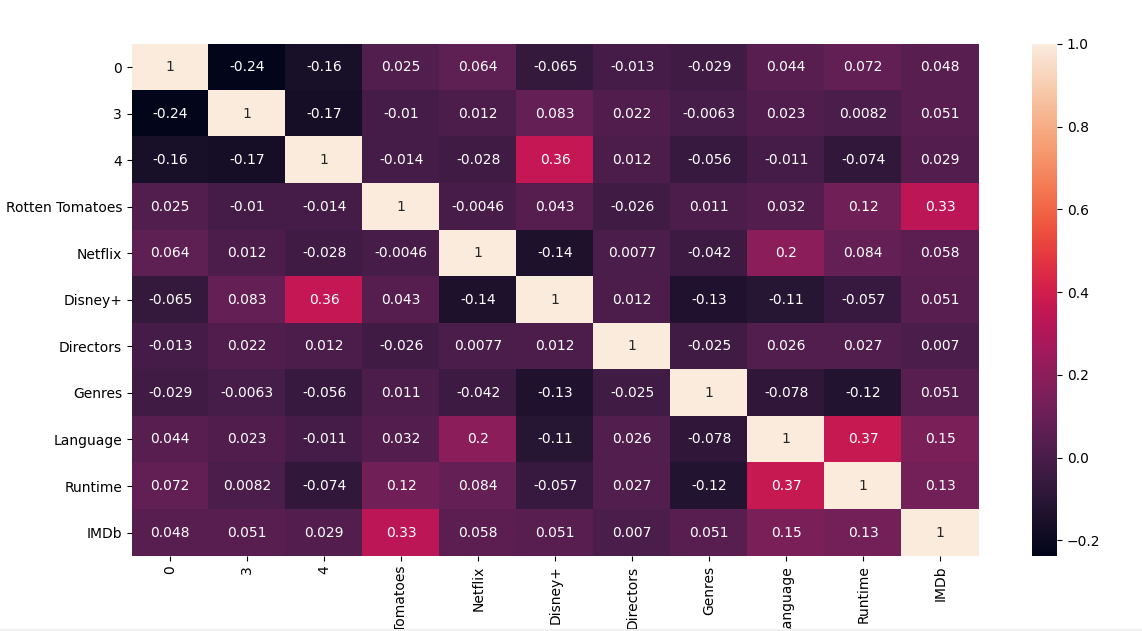
**For more Data (**flow of code & flow chart**).**

Flow of code:

1. Drop 1 columns that have the most missing values and the title (doesn't affect the result).

* According correlation

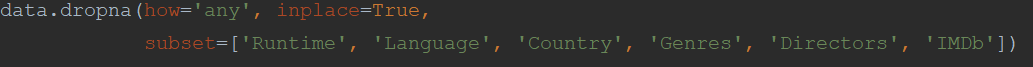




* Drop (title).

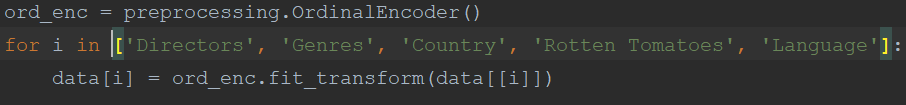


* Drop the rows that contain missing values (value = NULL).

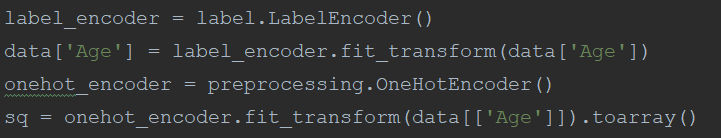


1. Map string values for numbers.

* preprocessing OrdinalEncoder:



* Preprocessing OneHotEncoder



1. Normalize data (fit data to be in the same scale).



1. Divide data into input and output.



1. Split data set into training and testing.



1. Regression:
2. Multiple Linear Regression

* Mean Square error = 0.019696934664050706

1. Polynomial Regression

* Mean Square error = 0.017397115586387072

Flowchart:

