1) **preprocessing Techniques**

- Acquire the dataset:
  To build and develop Machine Learning models, you must first acquire the relevant dataset. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. For instance, a business dataset will be entirely different from a medical dataset. While a business dataset will contain relevant industry and business data, a medical dataset will include healthcare-related data.

- Import all the crucial libraries

  we'll show you how to import Python libraries for data preprocessing in Machine Learning:
  Numpy, Pandas, Matplotlib, sklearn
  import numpy as np
  import pandas as pd
  from sklearn.preprocessing import PolynomialFeatures, LabelEncoder

- Import the Dataset
  Data = pd.read_csv('movies_training.csv')

- Extract the independent variables
  X = data.iloc[:,1:12]
  Y = data['IMDb']

- Identifying and handling the missing values

  1- Drop Columns and Rows:

  data.dropna(axis = 0 ,how='any', inplace=True)

  data.drop('Age',axis=1, inplace=True)

2- Calculating the mean

```
Z = Z.groupby(Z.columns, axis = 1).transform(lambda x:
x.fillna(x.mean()))
```

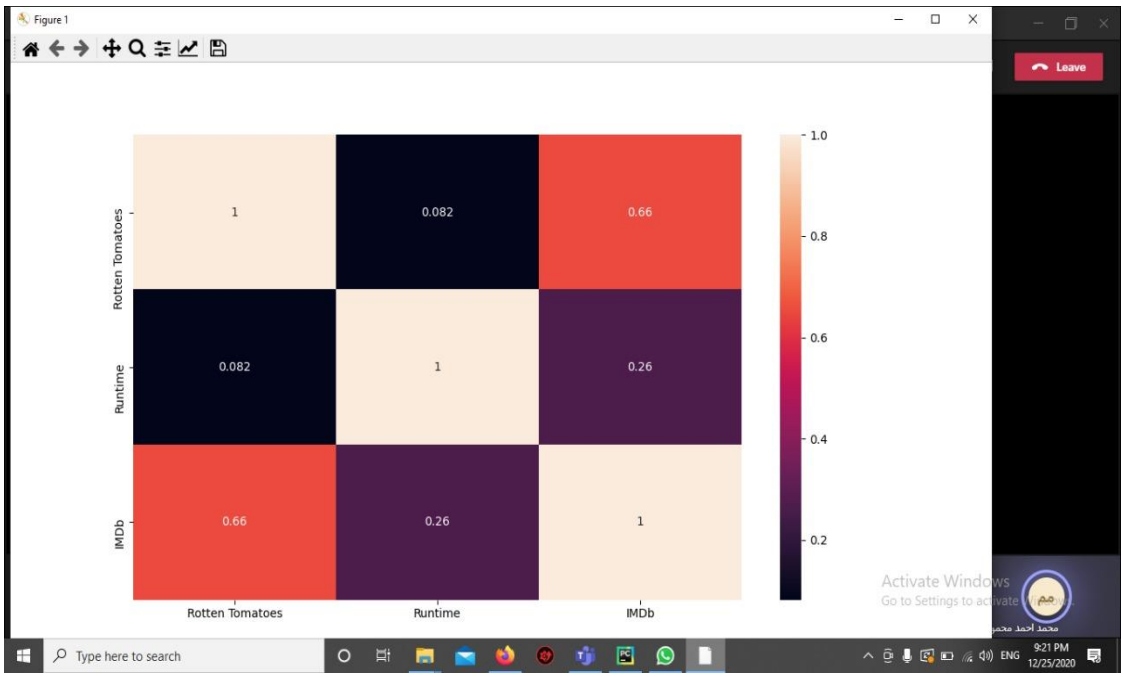
- Encoding the categorical data

```
for c in cols:
    lbl = LabelEncoder()
    lbl.fit(list(X[c].values))
    X[c] = lbl.transform(list(X[c].values))
```

- Hot Encoding

```
A= data['Title']
B= X['Genres']
# treat null values
X['Genres'].fillna(' ', inplace = True)
# separate all genres into one list, considering comma + space
as separators
genre = X['Genres'].str.split(',')
# flatten the list
flat_genre = [item for sublist in genre for item in sublist]
# convert to a set to make unique
set_genre = set(flat_genre)
# back to list
unique_genre = list(set_genre)
# create columns by each unique genre
Z= X.reindex(X.columns.tolist() + unique_genre, axis=1,
fill_value=0)
# for each value inside column, update the dummy
for index, row in Z.iterrows():
    for val in row.Genres.split(','):
        if val != ' ':
            Z.loc[index, val] = 1
Z.drop('Genres', axis = 1, inplace = True)
```

2)



After analysis become Top Features are Rotten Tomatoes and Run time, And Age column has low correlation and type column has only zeros in dataset so, I remove them

3) Regression Techniques

First Technique : we use multiple linear regression

multiple Linear regression is a basic and commonly used type of predictive analysis which usually works on continuous data. Multiple Linear regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables or Predictor variable and Target variable. It also assumes that there is no major correlation between the independent variables. Multi Linear regressions can be linear and nonlinear. It has one y and two or more x variables or one dependent variable and two or more independent variables.

Second Technique : we use polynomial regression

Polynomial Regression is a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an *nth* degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x). Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable.

4) difference between multiple linear regression and polynomial regression

there is no major correlation between the independent variables in multiple linear regression. if we know that our data is correlated, but the relationship doesn't look linear? So hence depending on what the data looks like, we can do a polynomial regression on the data to fit a polynomial equation to it.

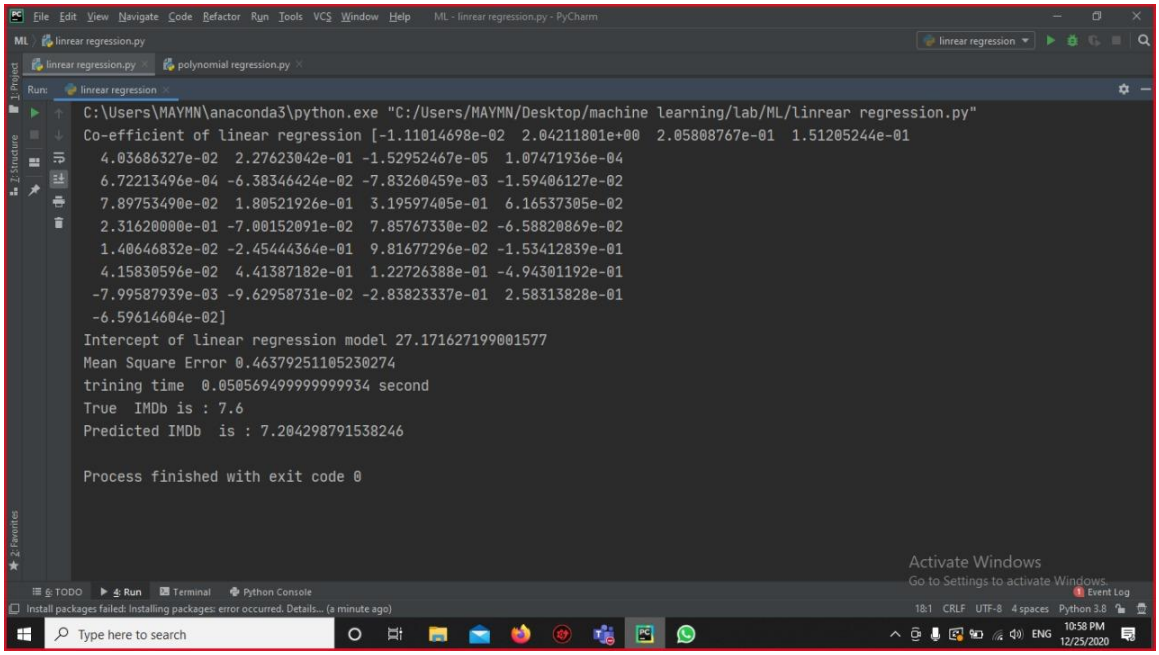
Result :

1) Multiple linear regression

   MSE : 0.43

   Intercept : 24.33

   training time : 0.05 second

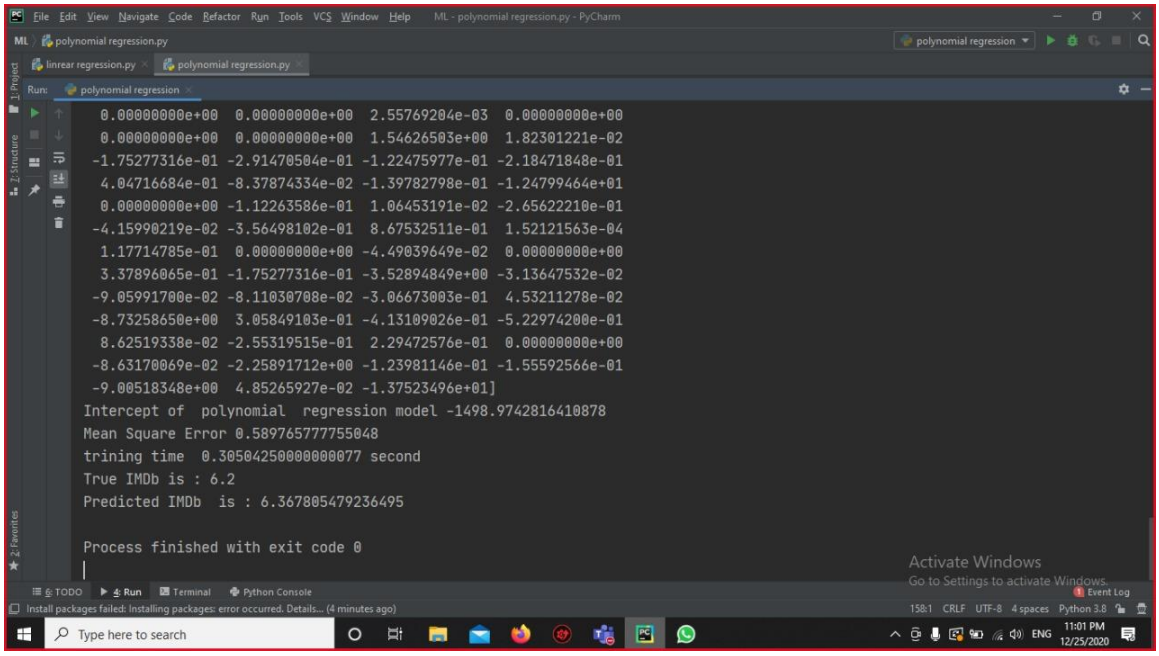


2) Polynomial regression

MSE : 0.55

Intercept : -1460.4

Training Time : 0.3



5) You must clearly mention what features you used or discarded to create your regression models.

(Year, Netflix , Hulu, Prime Video , Disney+ , Type , Directors , Genres Country , Language , Runtime)

6) sizes of your training, testing

Training : 70%
Testing: 30%

8) You should include screenshots of the resultant(s) regression line plots if possible or any data visualization.

9) at first we remove Rotten Tomatoes column and MSE > 1 but when we edit that column and add mean values in nan cells MSE < 1 almost 0.4