

1 MILESTONE 1

Predicting Song Popularity

1. Preprocessing techniques:

The used preprocessing technique is **Label Encoder** by sending the features list and test data to the function to be fitted then transformed to be ready to be used.

```
from sklearn.preprocessing import LabelEncoder

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

3. Regression techniques:

1. Linear Regression.
2. Polynomial Regression.

4. Analysis:

1. Linear Regression: 1. Training Time

```
Training time: 13.961836338043213s
```

2.Results

```
Co-efficient of linear regression [ 1.66636690e+00  6.64619462e-01 -3.91594514e+00 -5.05514084e-
 2.40219110e-01  1.06300151e-06 -2.21998676e+00 -6.36673592e-01
 1.79774962e-05 -5.73577509e-04 -1.49968680e-03 -3.86802793e+00
 9.68532647e-02 -6.00775953e-02 -4.61586186e-06 -2.73848432e-04
 3.22461254e-04]
Intercept of linear regression model -1278.6198010823507
Mean Square Error 122.95993691860593
True value for the first song in the test set is : 58
Predicted value for the first song in the test set is : 55.82727954703864
```

2. Polynomial Regression: 1.Training Time

```
Training time: 11.295044183731079s
Co-efficient of Polynomial regression
```

2.Results

```
1.20859769e-03  1.84780108e-08  2.86849648e-01  1.46377917e-01
6.35260359e-01  1.64120355e-05  5.95342963e-04 -2.58883756e-04
-6.77661318e-03  4.41661966e-02  7.67736849e-07  1.56447560e-03
8.71354016e-06  1.49673310e+01 -6.12694383e-07  7.61341666e-03
-4.85212674e-05  1.96749284e-10  1.38551632e-08  2.45799825e-10
1.46102320e-04  1.19171269e-06 -1.43386458e-09]
Intercept of linear regression model -1161.378884295635
Mean Square Error 116.00269514720523
True value for the first player in the test set : 35
Predicted value for the first player in the test set : 36.85578921891647
```



5. Features:

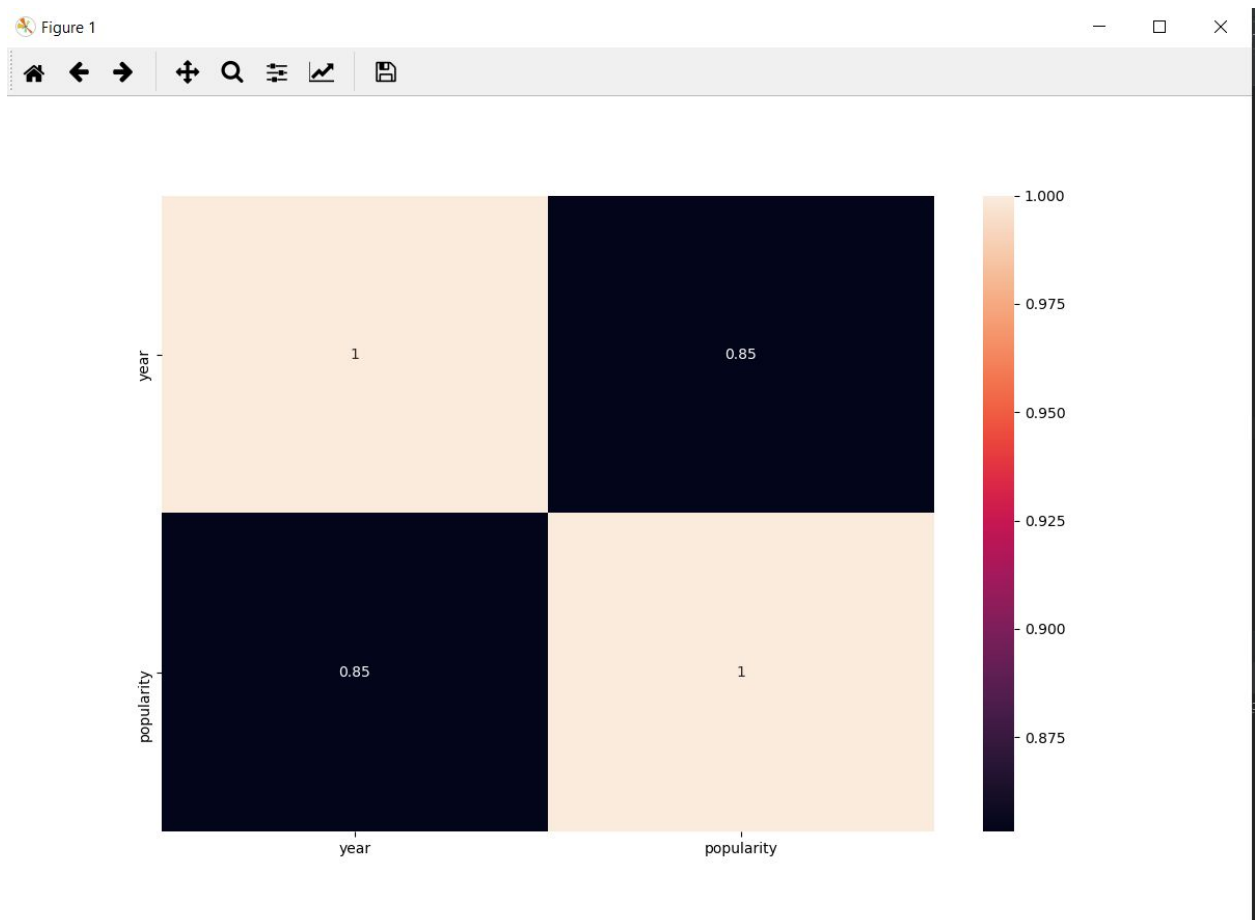
1. Linear Regression: Used All features except **all nulls** and '**explicit**' feature.
2. Polynomial Regression: Used all features except **all nulls**.

6. Size: 1. Linear Regression: **Training set = 70% of the dataset**
Testing set = 30% of the dataset

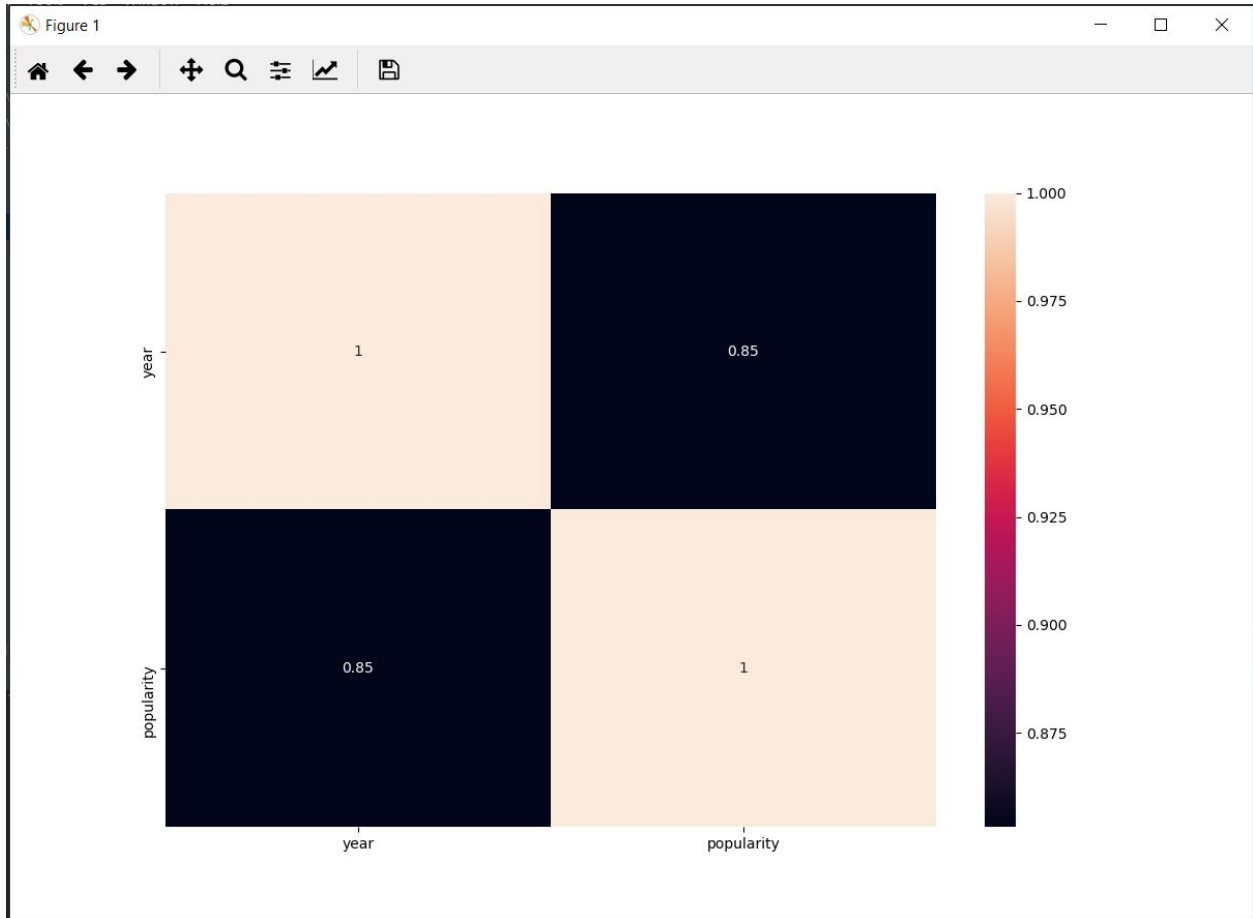
2. Polynomial Regression: **Training set = 70% of the dataset**
Testing set = 30% of the dataset

7. Data visualization:

1. Linear Regression:



2. Polynomial Regression :





8. Conclusion

This model needs more trials and testing to get better values because as we saw the polynomial regression got better results and lesser MSE than linear regression even after decreasing feature numbers in the linear model and we need to try another features set to determine whether the existing one is doing poorly or not.

■ ■ ■