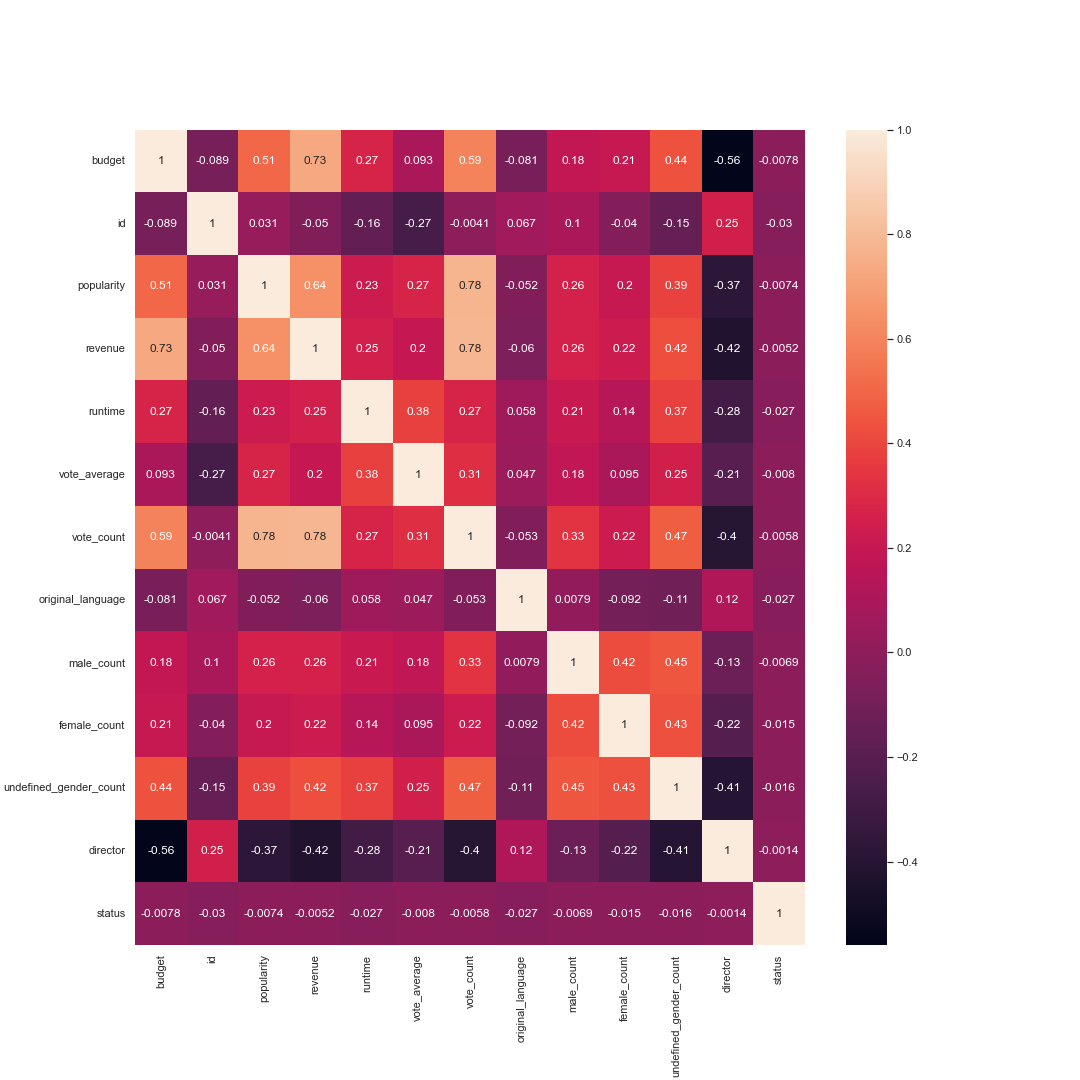
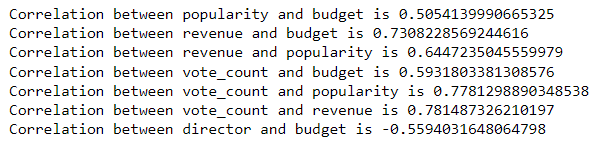
**Data set 1 – 46H**

**Heat Map**:- Correlation between most of the features is very low i.e, data redundancy is low.



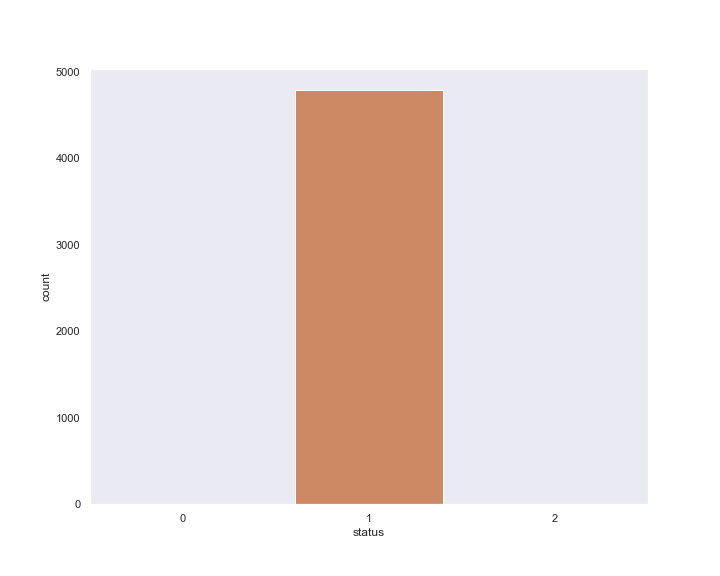
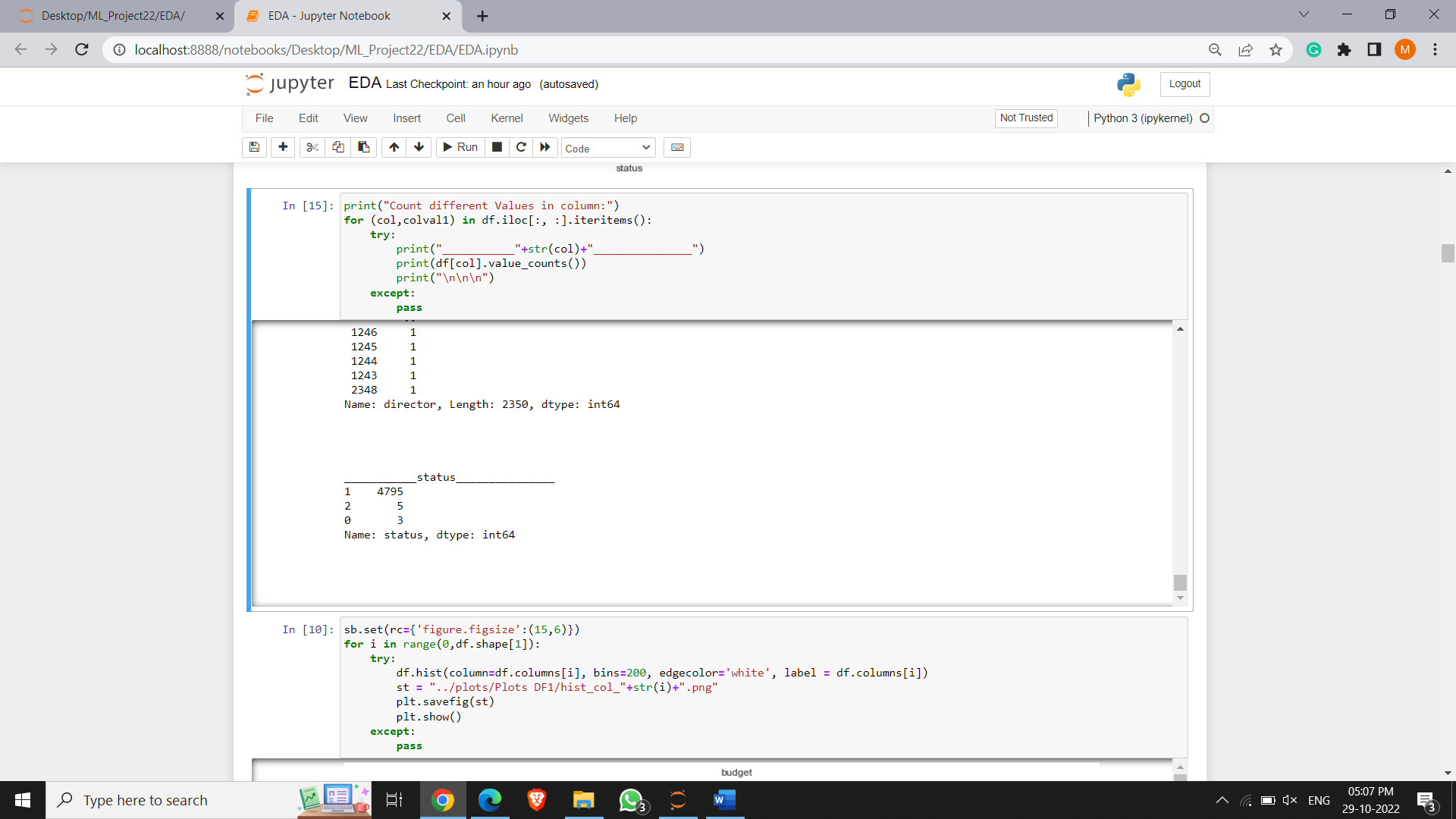
A very few Features in the Dataset are Highly Correlated.

Highly Correlated Features and their Correlations,

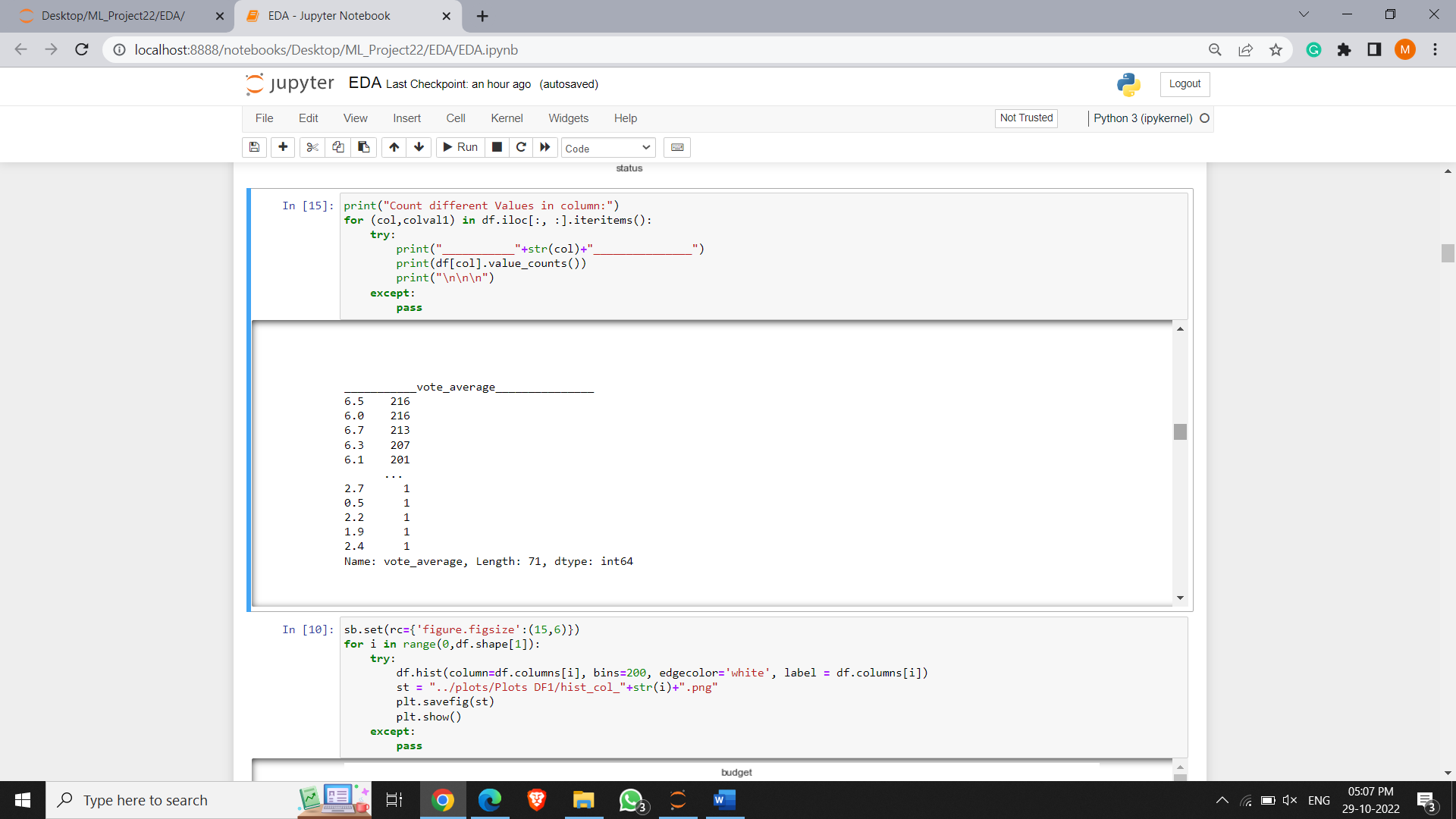
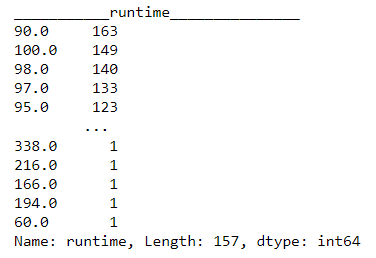
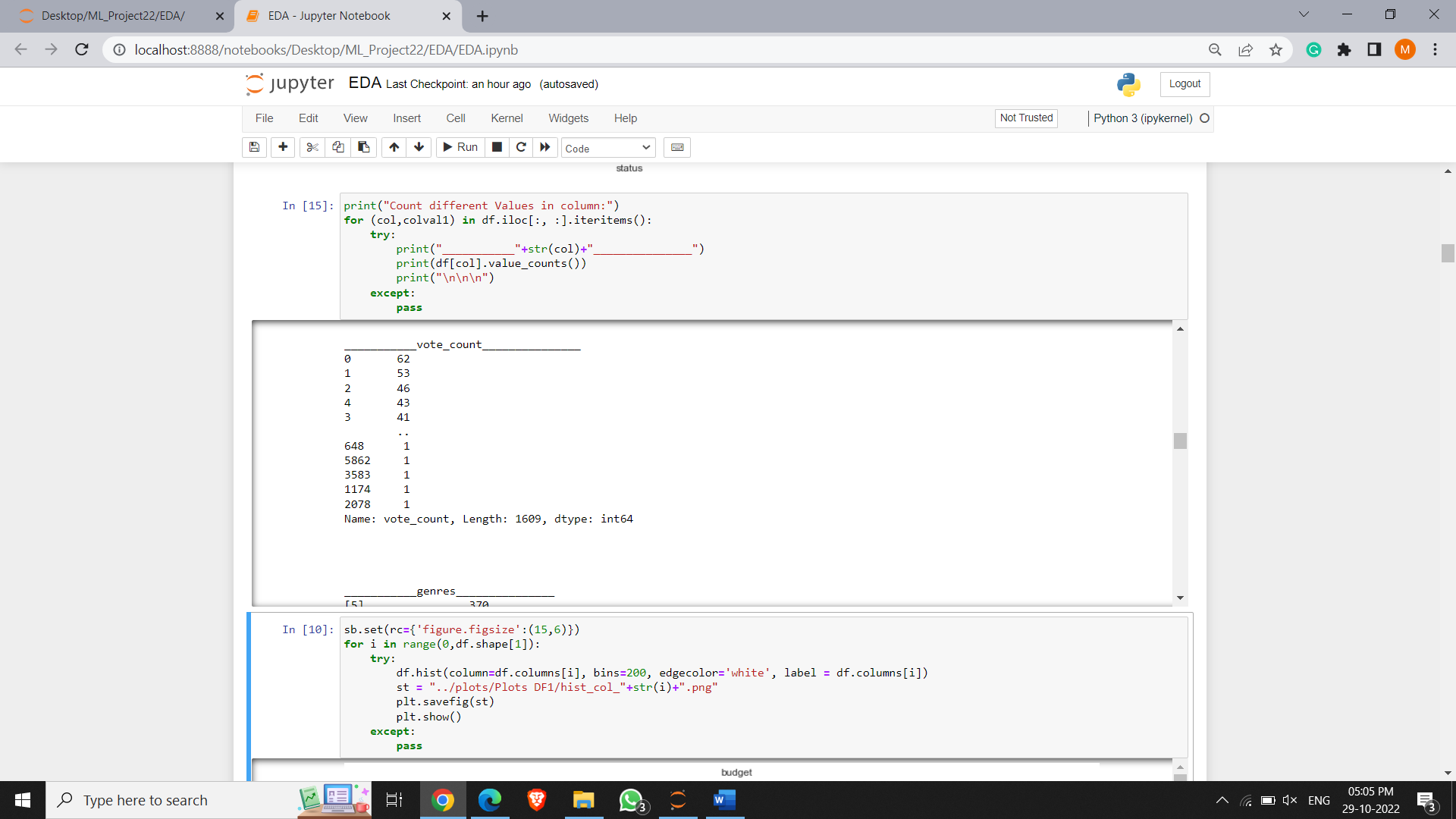


**Data Distribution-**

Data is not equally distributed for feature status.

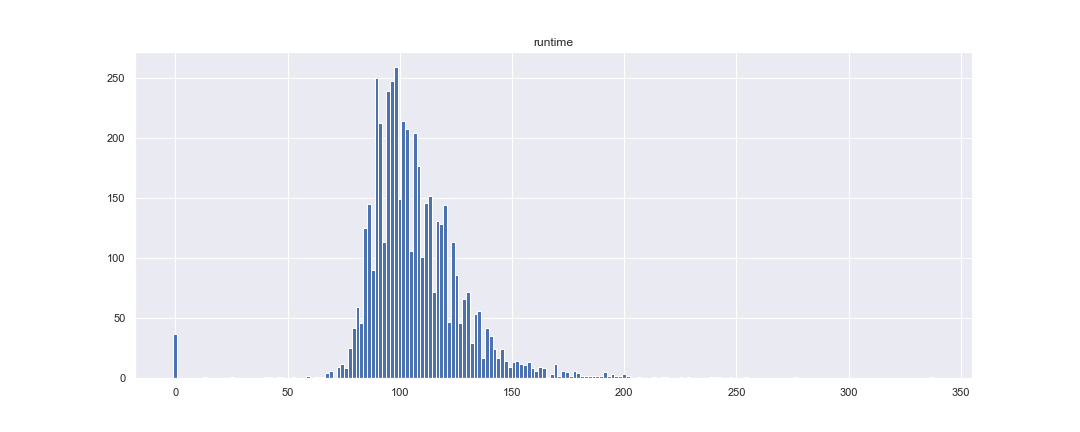
 

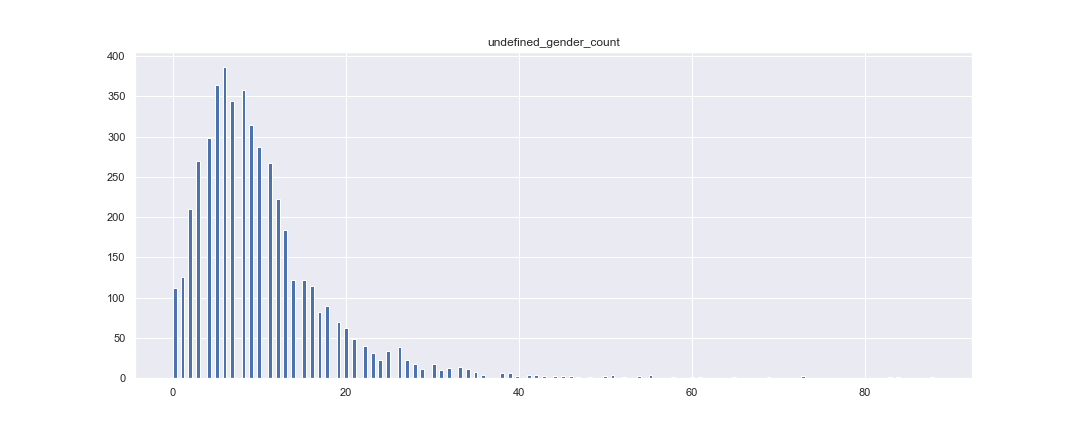
For the remaining features, Data is fairly distributed.

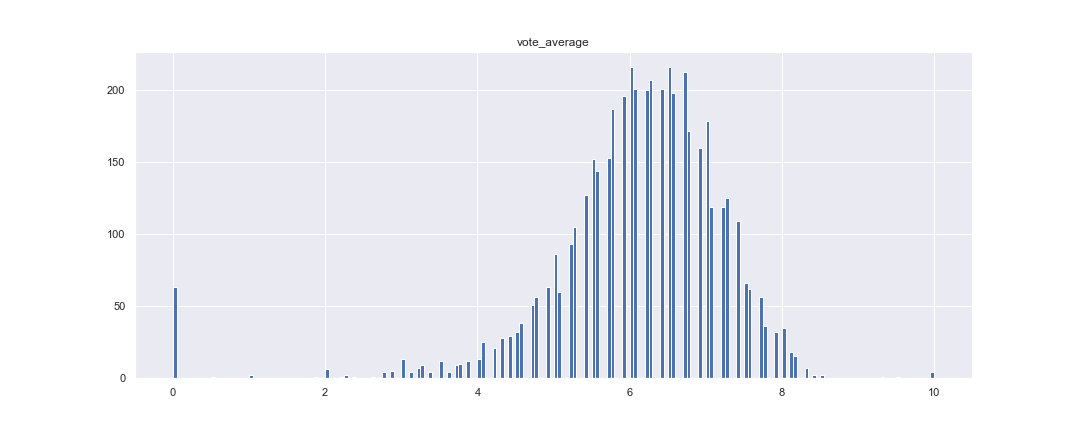


**Histogram –**

Only a selected number of features are normally distributed ( forms a bell-shaped curve).





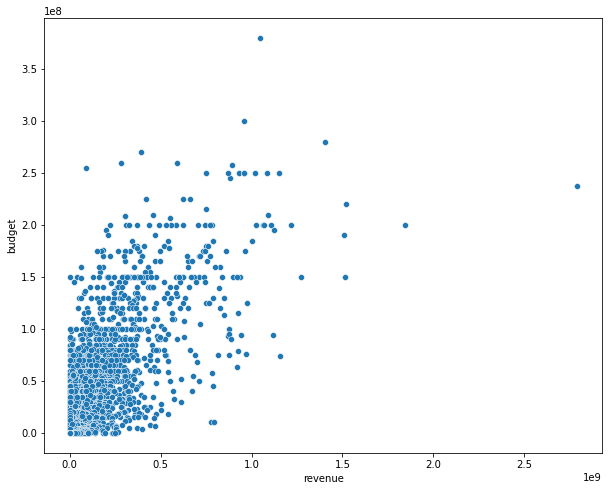


**Scatter Plot –**

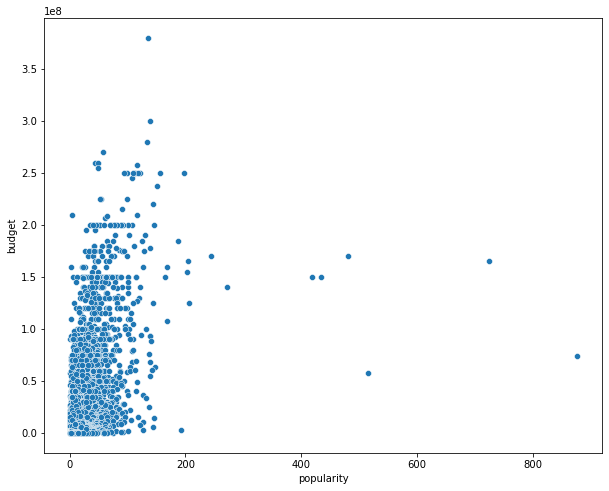
Scatter Plots of some of the highly correlated features –

Popularity and revenue – Correlation

Correlation between revenue and budget is 0.7308228569244616



Correlation between popularity and budget is 0.5054139990665325

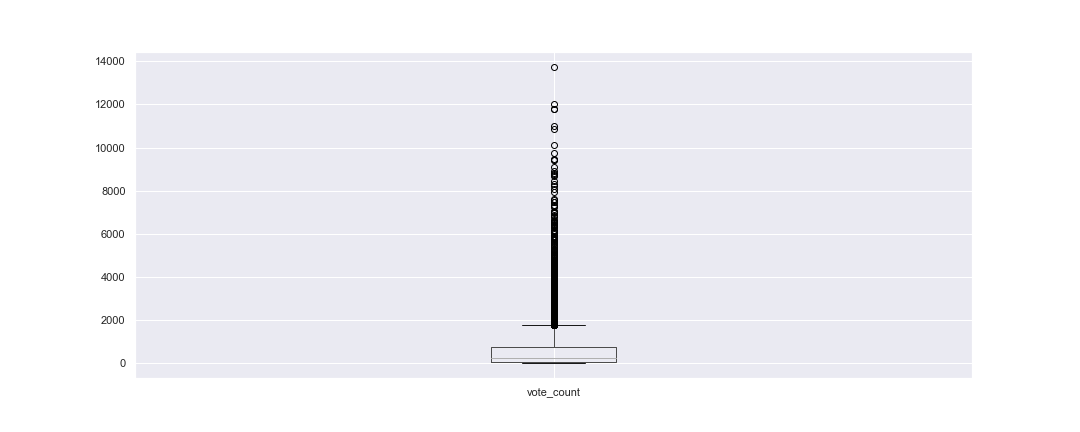
****

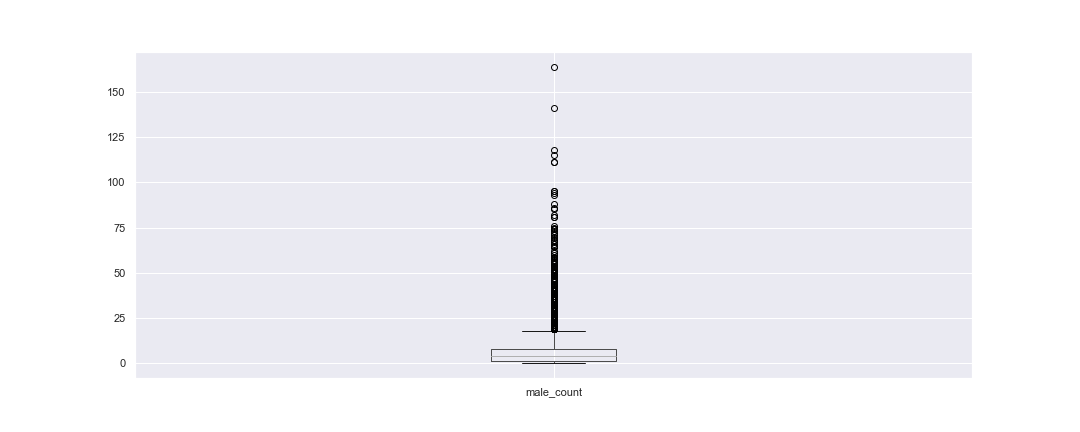
While the correlation between this features is high no clear relation can be established from the scatter Plots

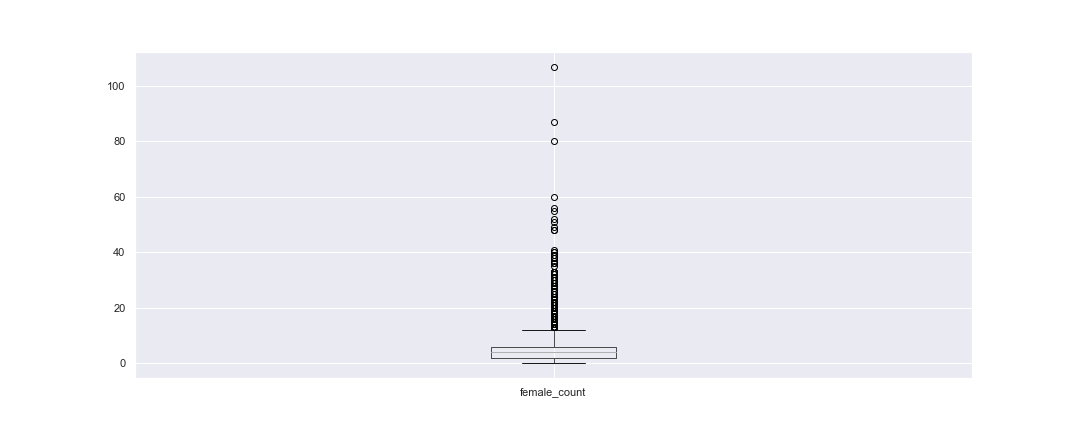
**Box Plot –**

From the Box Plots, all features contain outliers.

For a few columns, the count and values of the outliers are way too high.





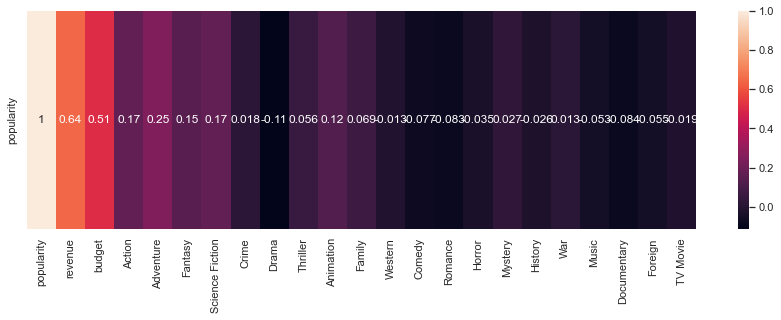


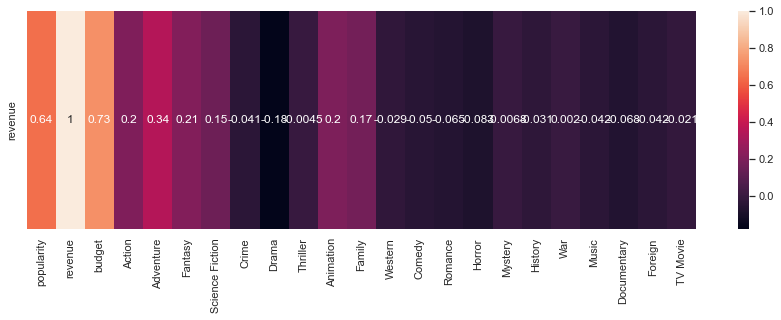
**Heat Map of Genres and Popularity/Revenue –**

From the Box Plot, we conclude that data contains too many outliers in almost all features.

The correlation between the Genre type and popularity is very low.

None of the genre types is correlated with Popularity.

****

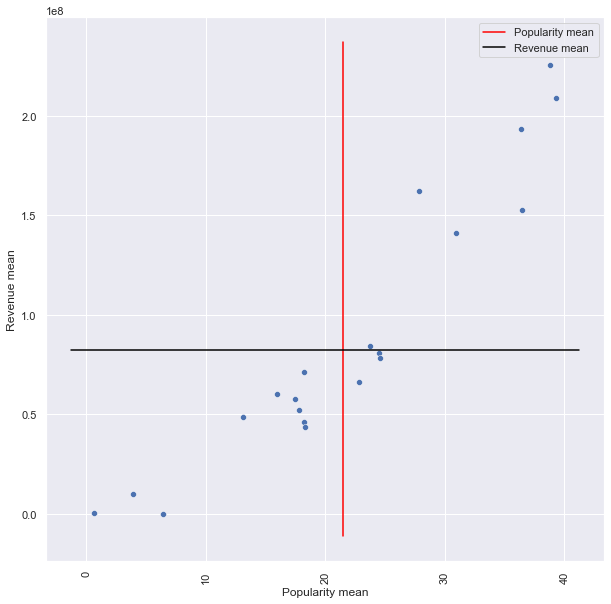
****

Similar results are observed with Revenue and Genre Types.

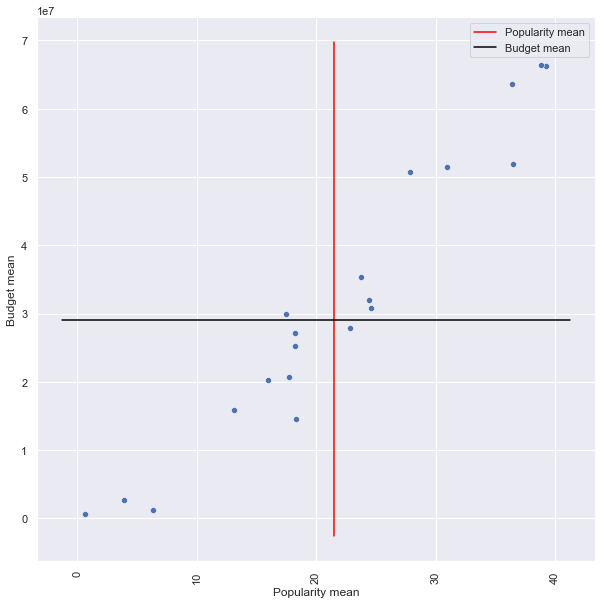
**Mean Popularity/Revenue/Budget with different Genre Types –**

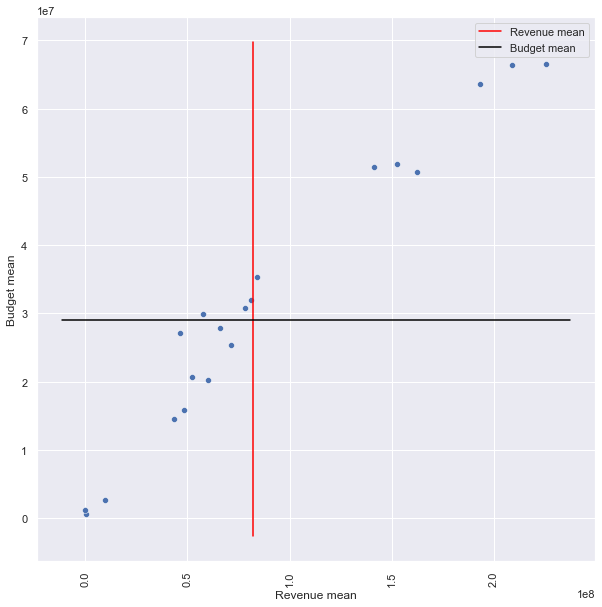
The mean values of the Genre Types individually are way off the mean value for the complete dataset.

****

****

Here Dots are the Different Genre Type’s means, the red line denotes the Popularity means of the complete dataset, and the Black line denotes the Revenue means of the complete dataset.

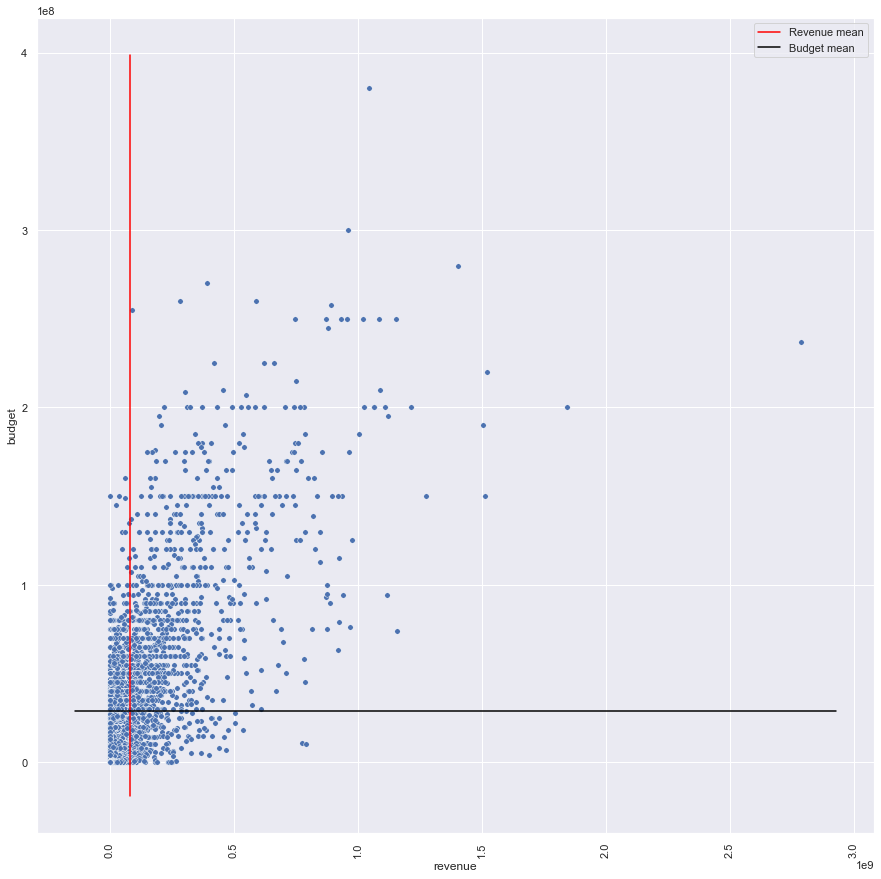
****

****

Similar results are observed between Popularity means and Budget means, and revenue means and budget means.

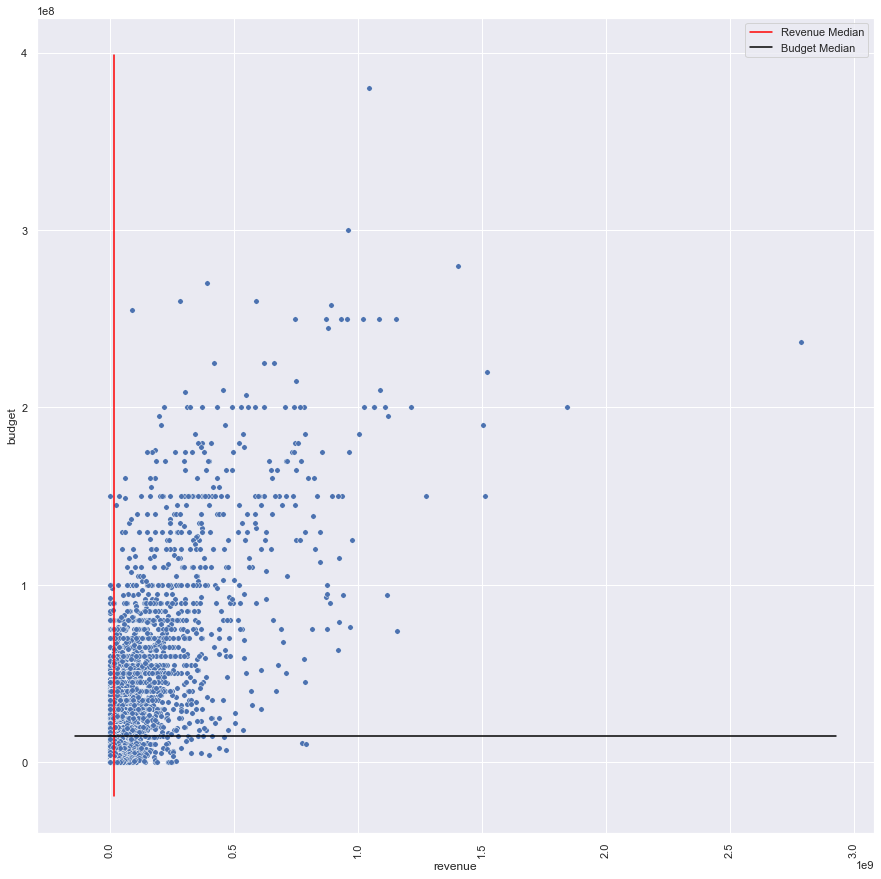
**Mean/Median Plots of the Dataset for Revenue and Budget with all data points -**

The mean of the Budget and Revenue lies somewhere in half of the data points.

****

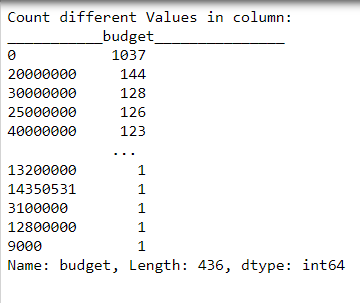
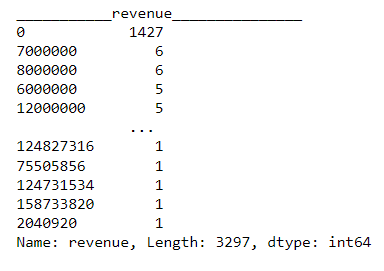
Mean lines lies somewhere in between all the datapoints

Whereas the median of the dataset for budget and revenue is very low compared to the values of the data points.

****

Most of the values lies above median

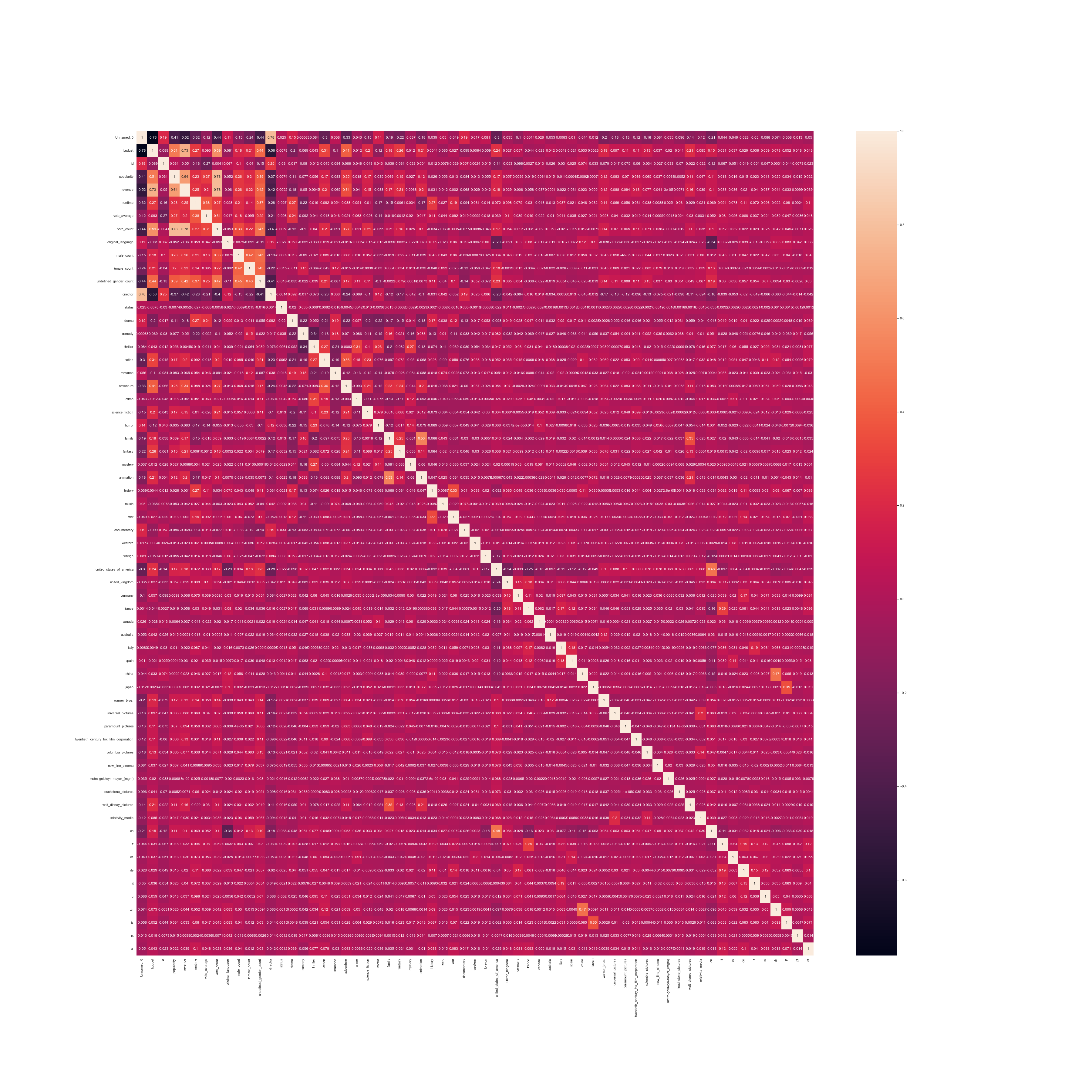
This could be due to the high count of zero values for the budget and revenue features in the dataset.

****

High Count of zero values

**Data set 1 – 46H – one hot encoded**

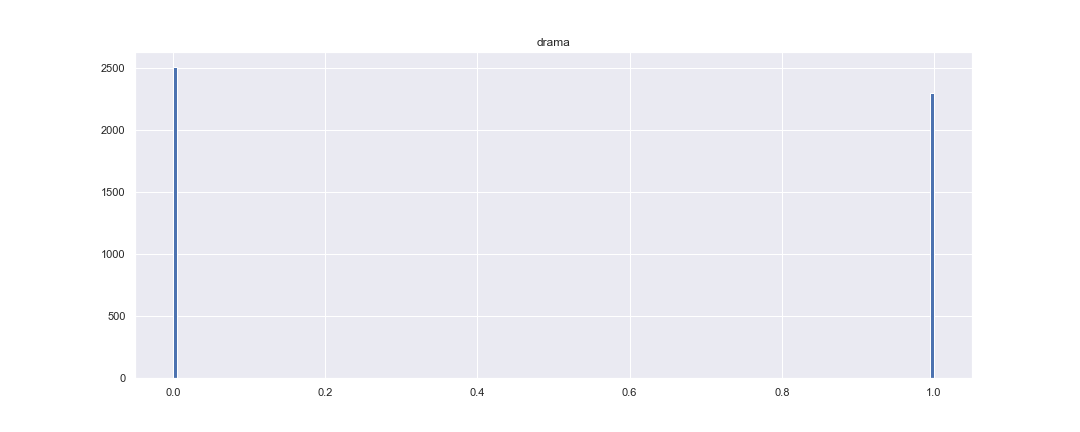
**HeatMap –**

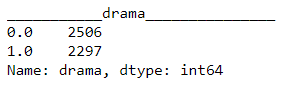


**Data Distribution –**

In the One hot encoded dataset the data is not fairly distributed amongst the top classes for the genre, keywords etc.

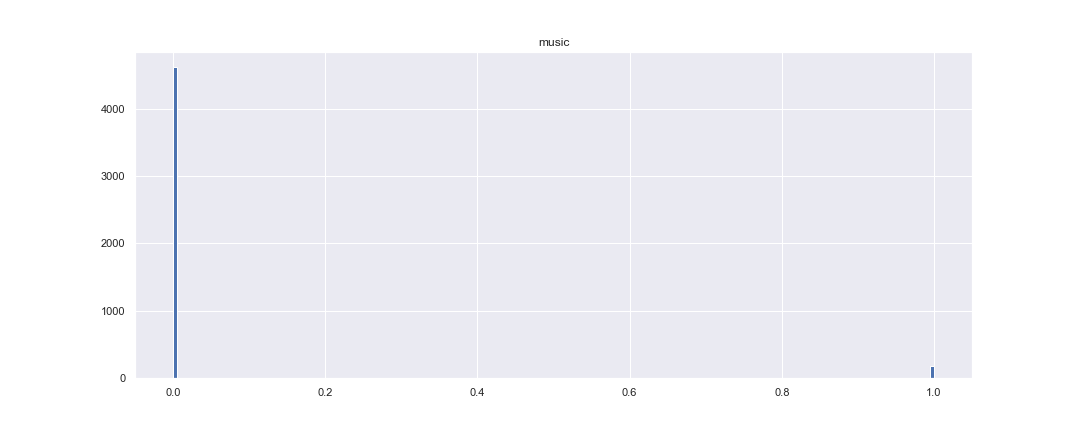
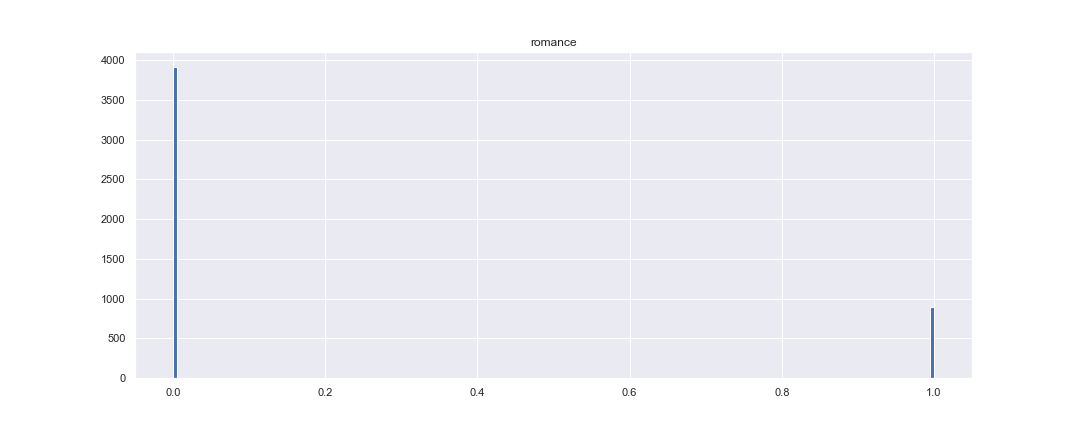
For a few features, the distribution is almost half.



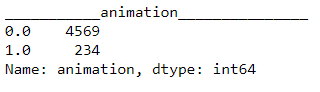
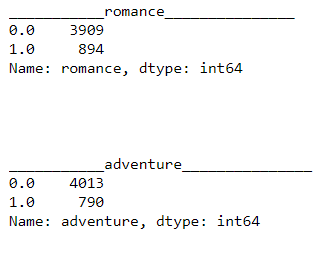


A fair distribution

Half, Whereas for most classes, the data is favourable for one type only.

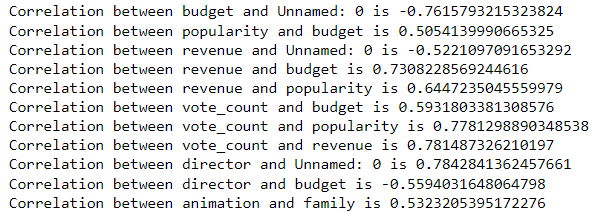


Not fairly Distributed

****

This is because we are only checking the top ten unique values of the features. When considering all the top ten collectively, the distribution of the dataset is very good.

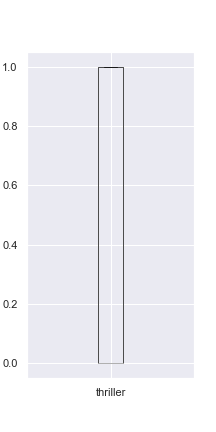
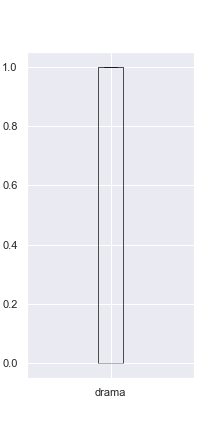
**Correlation among the Classes –**

****

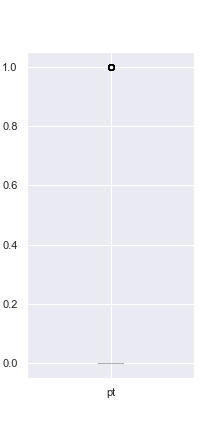
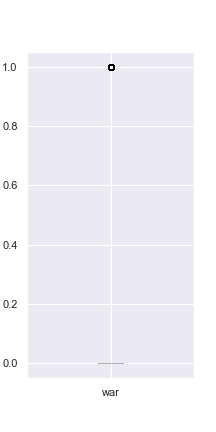
Amongst all different, only two features, namely animation and family, are highly correlated with a correlation value of 0.53.

**Bar Graph –**

For only few out of all the classes the the mean lies near the five.



Equal Distribution



Most values lies at zero