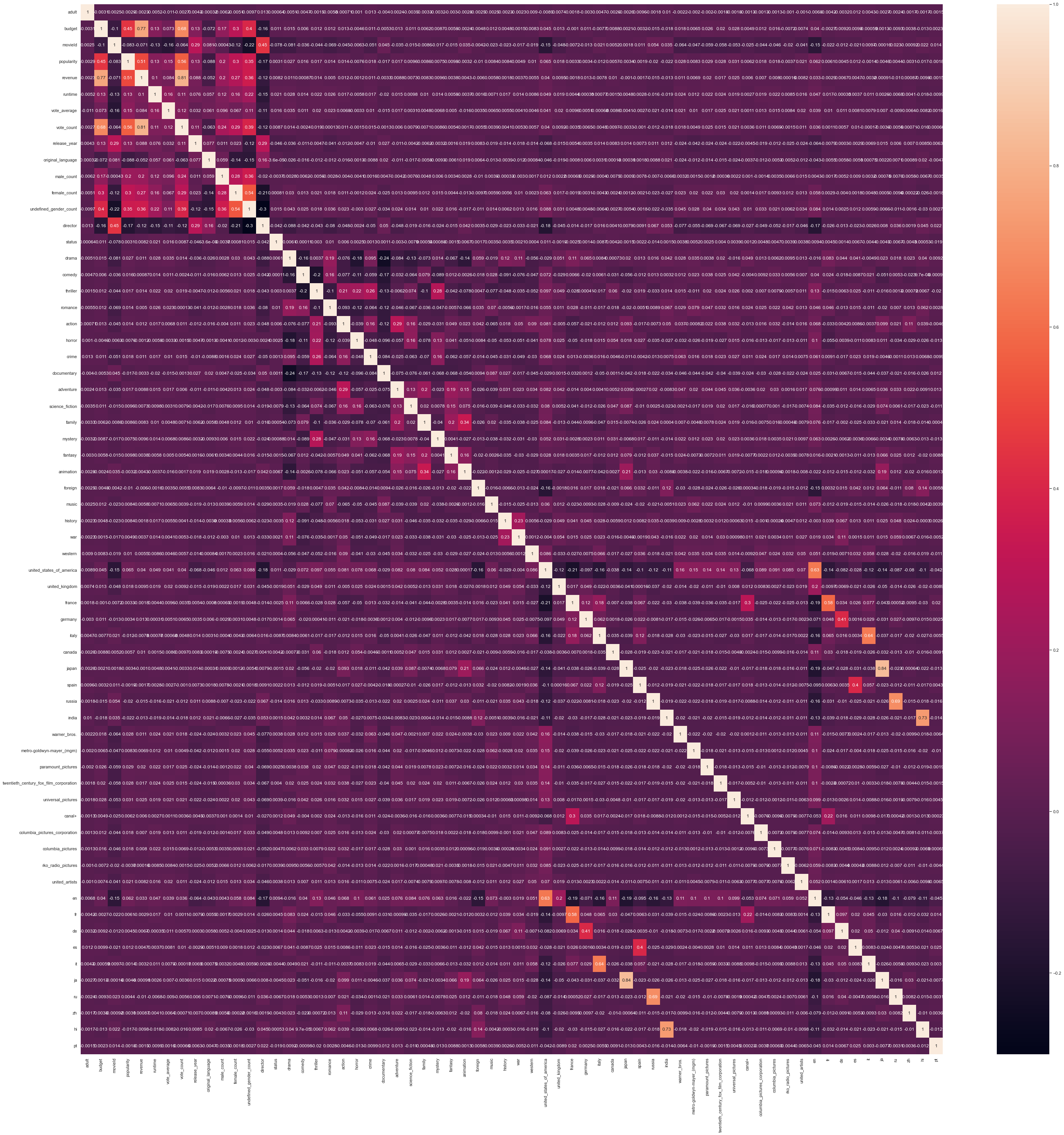
**Data set 1 – 46k**

**Shape –**

(46195, 70)

**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,

Correlation between revenue and budget is 0.7686798095644242

Correlation between popularity and revenue is 0.5057940307296539

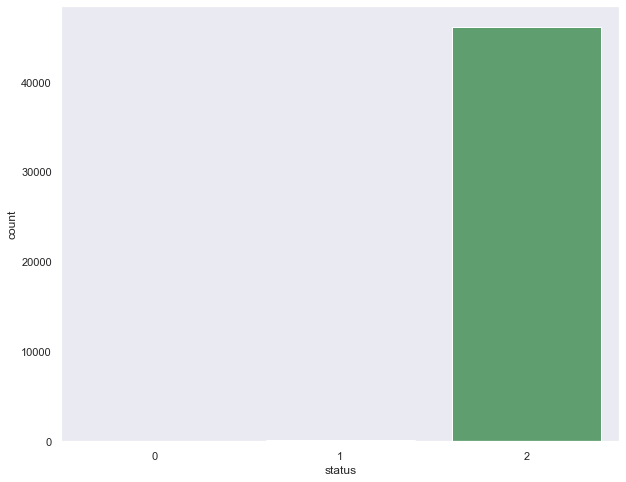
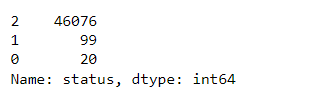
Correlation between vote\_count and budget is 0.6765841628857977

Correlation between vote\_count and popularity is 0.5604841570229393

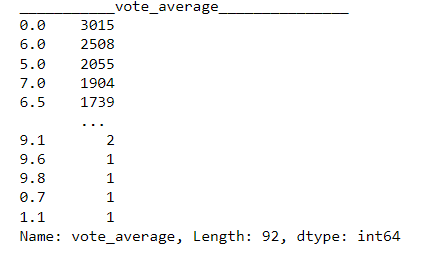
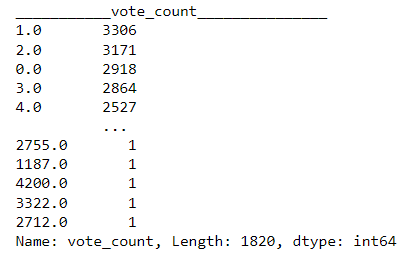
Correlation between vote\_count and revenue is 0.8119905411390768

**Data Distribution-**

Data is not equally distributed for feature status.

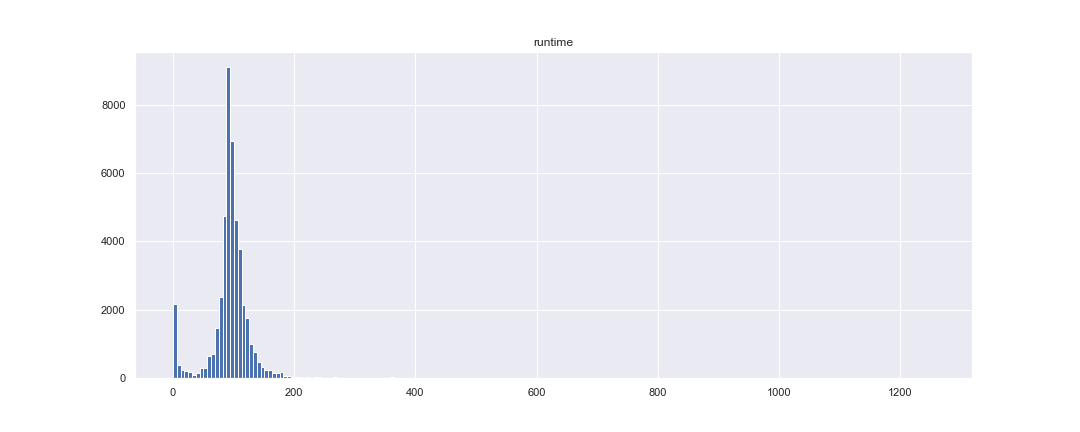
 

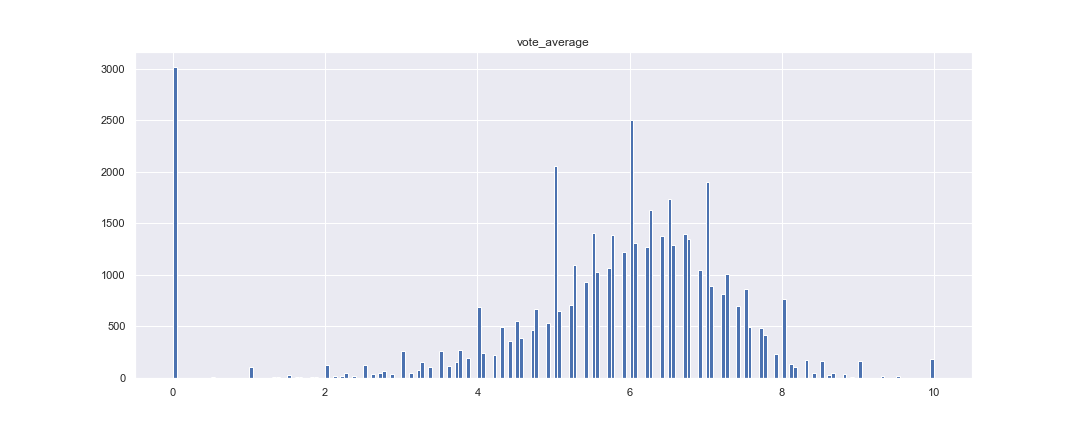
For the remaining features, Data is fairly distributed.

**Histogram –**

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





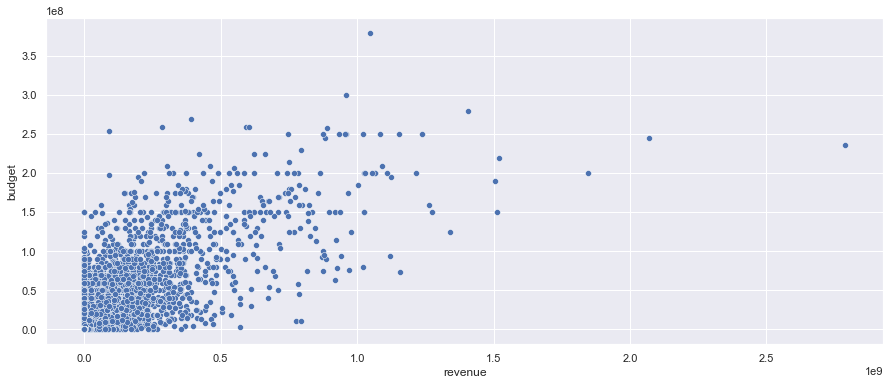
Larger number of values in voting average is zero

**Scatter Plot –**

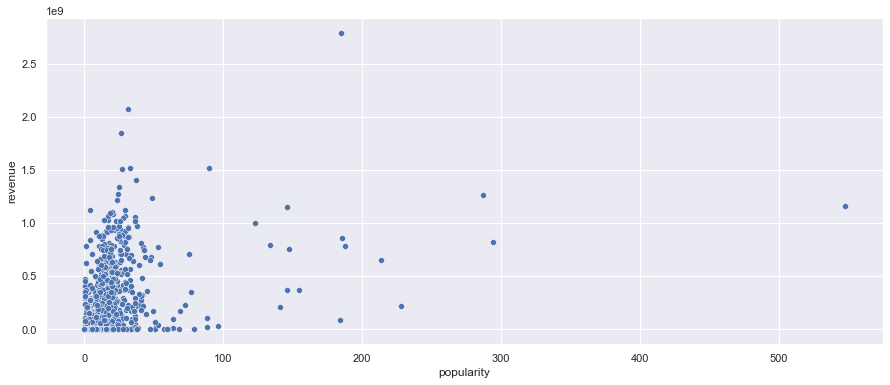
Scatter Plots of some of the highly correlated features –

Popularity and revenue – Correlation

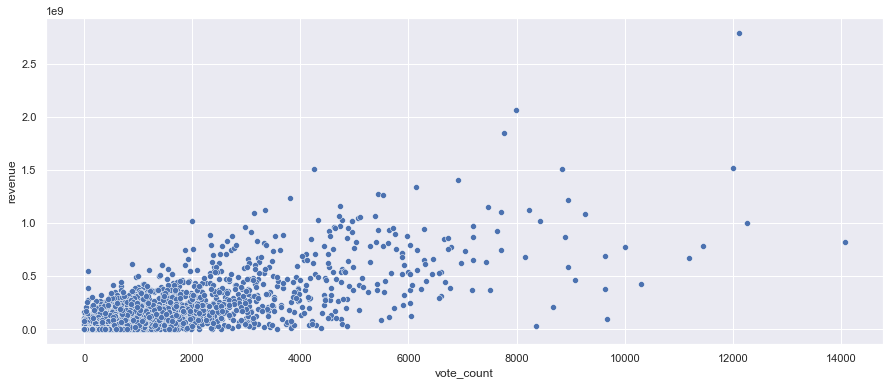
Correlation between revenue and budget is 0.7686798095644242



Correlation between popularity and revenue is 0.5057940307296539

****

Correlation between vote\_count and revenue is 0.8119905411390768

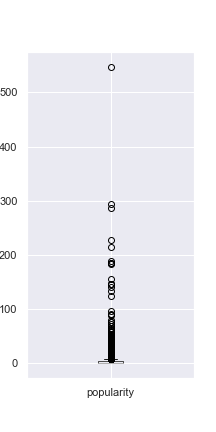
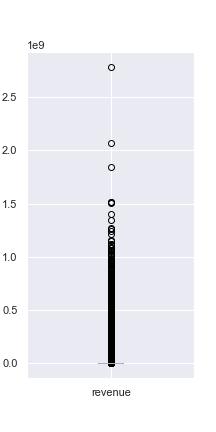
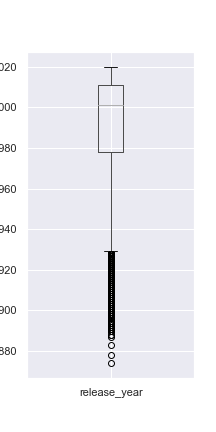
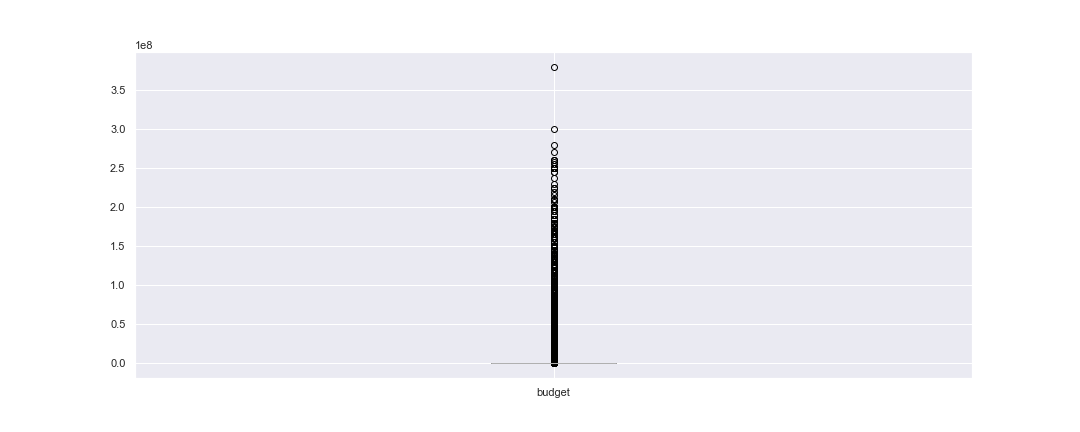


While the correlation between this features is high no clear relation can be established from the scatter Plots

**Box Plot –**

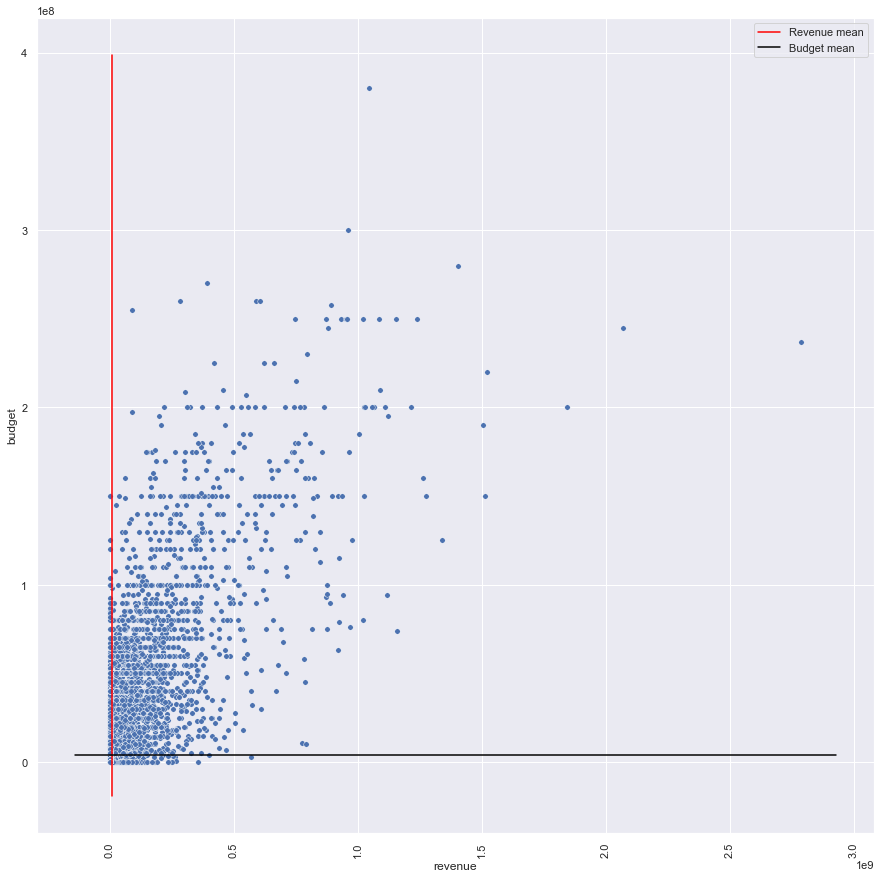
From the Box Plots, all features contain outliers.

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



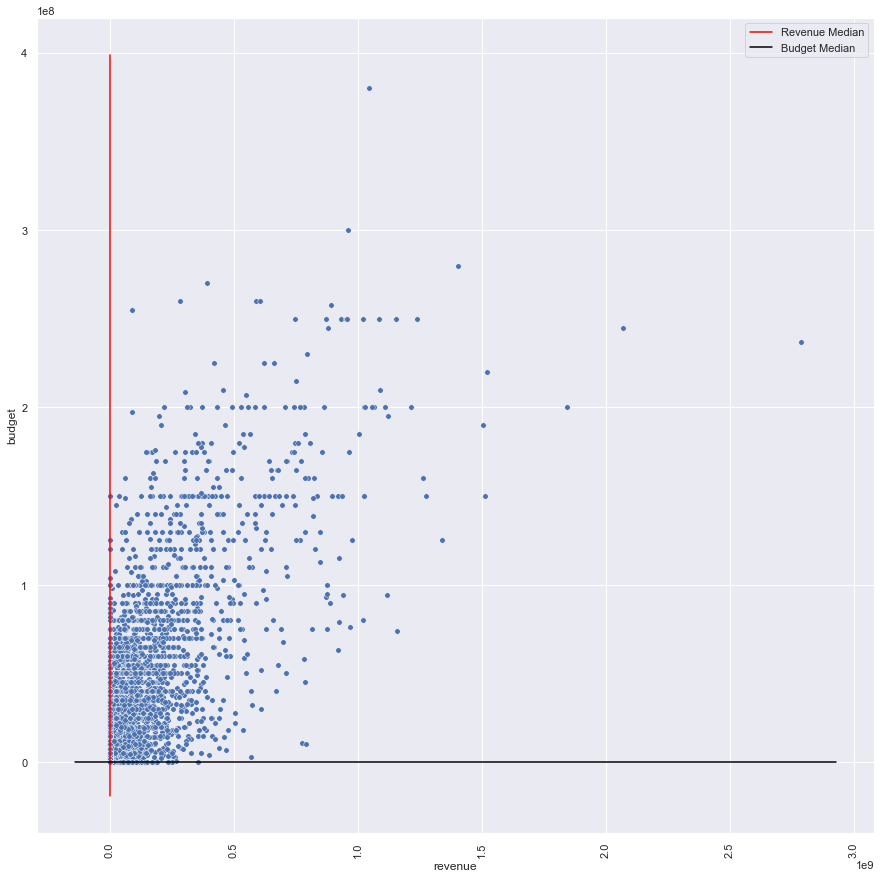
**Mean/Median Plots of the Dataset for Revenue and Budget with all data points -**

The mean of the Budget and Revenue is very low as compared to the data points.

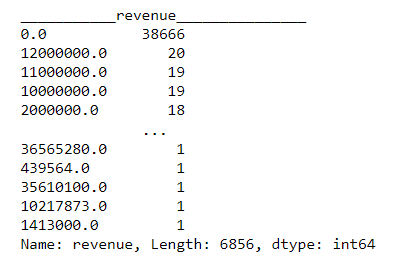
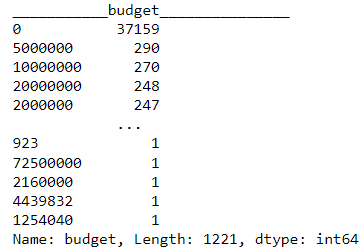
****

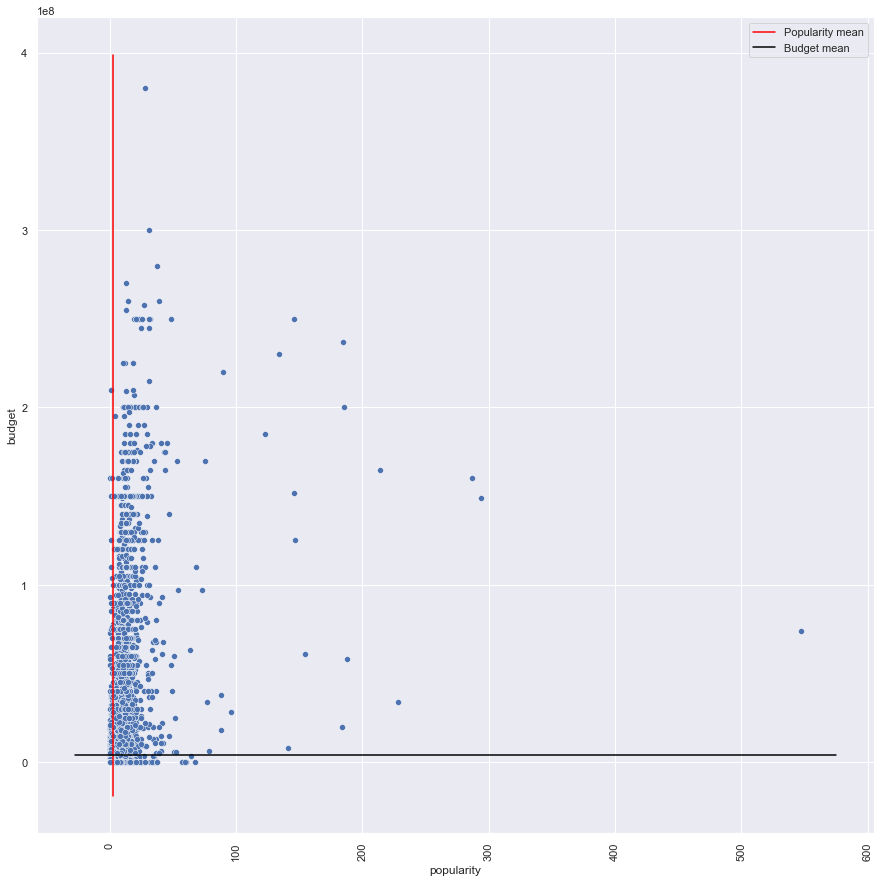
Here Dots are the datapoints, the red line denotes the Revenue means of the complete dataset, and the Black line denotes the Budget means of the complete dataset

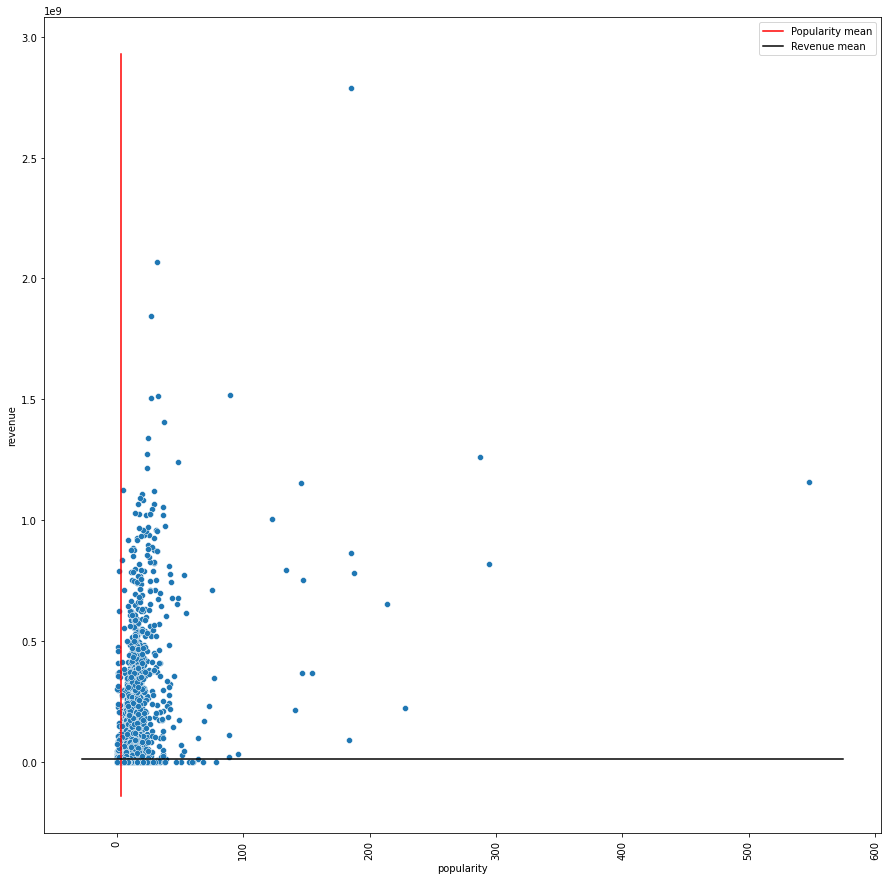
The median of the dataset for budget and revenue is very low compared to the values of the data points.

****

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

****

****

****

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